whether or how a given algorithmic behavior should be addressed.

If a representation is factually accurate, can it still be algorithmic unfairness?
Yes. For example, imagine that a Google image query for "CEOs" shows predominantly men. Even if it were a factually accurate representation of the world, it would be algorithmic unfairness because it would reinforce a stereotype about the role of women in leadership positions. However, factual accuracy may affect product policy's position on whether or how it should be addressed. In some cases, it may be appropriate to take no action if the system accurately affects current reality, while in other cases it may be desirable to consider how we might help society reach a more fair and equitable state, via either product intervention or broader corporate social responsibility efforts.

If a system's behavior is not intended, can it still be algorithmic unfairness?
Yes. If the behavior is unfair, it meets the definition regardless of the root cause.

Definition

"algorithmic unfairness" means unjust or prejudicial treatment of people that is related to sensitive characteristics such as race, income, sexual orientation, or gender,[5] through algorithmic systems or algorithmically aided decision-making.

More on misinformation and our role within it.

---------- Forwarded message ----------
From: [redacted]
Date: Tue, Nov 28, 2017 at 12:27 AM
Subject: Fake News-letter 11/27: Efforts to combat spread of (mis/dis)information
[redacted]

Fake News-letter

BY gTech Search + Content [redacted]

NOVEMBER 27, 2017 | INTERNAL ONLY

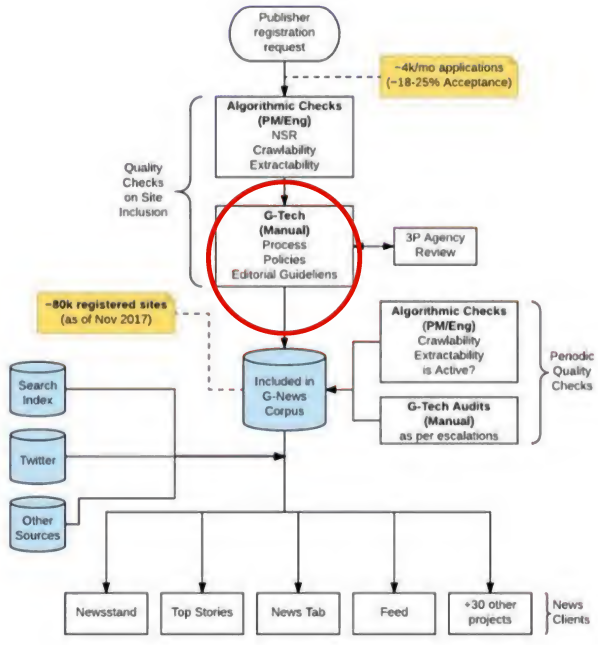
SOURCE QUALITY

Revamping News Corpus

Goal: Establish "single point of truth" for definition of "news" across Google products. Mitigate risk of low-quality sources and misinformation in Google News corpus.
Status: Define new Google News corpus utilizing existing infrastructure and tools while integrating different quality tiers and labels, algo-human content review, new cross-product review/exclusion pipeline and updated inclusion UI.

CONTACTS: [redacted]
MORE INFO: Proposal, Updated rater template (WIP) / Stress Test publisher list
NEXT STEPS: Research and gather info about possible signals to leverage for product teams to make source quality decisions. Finalize and test new rater inclusion guidelines with standards to combat misinformation.

News Ecosystem



Definition of Algorithmic Unfairness

[redacted]

Last Updated
February 2017

go/algorithmic-unfairness-definition

Goals

Our goal is to create a company-wide definition of algorithmic unfairness that:

1. **Articulates the full range of algorithmic unfairness that can occur in products.** This definition should be robust across products and organizational functions.
2. **Establishes a shared understanding of algorithmic unfairness** for use in the development of measurement tools, product policy, incident response, and other internal functions.[1]
3. **Is broadly consistent with external usage of the concept.** While it is not a goal at this time to release this definition externally, it should represent external concerns so that we can ensure our internal functions address these concerns.[2,3,4]

Non-Goals

The following are not goals for this document:

1. **Specify whether and how Google will take action** on potential instances of unfairness that involve the use of an algorithm. This will fall instead to product policy.
2. **Describe the consequences of algorithmic unfairness and why they matter**.

Definition

"algorithmic unfairness" means unjust or prejudicial treatment of people that is related to sensitive characteristics such as race, income, sexual orientation, or gender,[5] through algorithmic systems or algorithmically aided decision-making.

---

[1] The discussion in Lipton (2016) suggests that instances that present as identical product behavior may upon investigation be revealed to have substantially different root causes and therefore require different remediations. Because the nature of an instance may not be evident *a priori*, a wide range of instances are likely to be reported and investigated together. (Zachary Chase Lipton. The Deception of Supervised Learning. KDnuggets News 16:n33, September 2016.)

[2] Peter Swire. Lessons from Fair Lending Law for Fair Marketing and Big Data. Future of Privacy Forum, and presented before the Federal Trade Commission Workshop on "Big Data: A Tool for Inclusion or Exclusion?" (2014).

[3] Solon Barocas and Andrew D. Selbst. Big Data's Disparate Impact. 104 California Law Review 671 (2016).

[4] ACM US Public Policy Council. Statement on Algorithmic Transparency and Accountability. January 12, 2017.

[5] This list is not intended to be exhaustive, and additional examples of sensitive characteristics appear in the FAQ.

Scope

The following are in scope for algorithmic unfairness:

In algorithms that drive **predictive** systems (e.g., personalization or device financing), this definition encompasses automated actions significantly adverse to the interests of a user or group of users,[6] on the basis of a characteristic that is sensitive in the context of a particular interaction.

In algorithms that drive **representational** systems (e.g., search), this definition encompasses the implied endorsement of content likely to shock, offend, or upset users sharing a sensitive characteristic, or to reinforce social biases.[7]

The following are *not* in scope for this definition but may be covered by other internal definitions and policies:

1. **Biased content**, for example user-generated content that appears in products (while the content itself is not in scope, the algorithmic handling of such content may be in scope)
2. **Insensitive designs** which occasion unfairness, and/or interface designs that are less usable by groups with sensitive characteristics
3. **Google's internal decisions** such as hiring or compensation, which may be influenced by algorithms

Test Cases for Scope

The following are illustrative examples of what is (and is not) in scope for algorithmic unfairness. Of the examples that are in scope, a subset may be determined by product policy to require remediation.

Google Display Ads for High-Paying Jobs

CMU published a study in 2015 with experiment-based observations, arguing that Google's ad serving system perpetuates gender bias on the basis of two campaigns that were found to target high salary jobs at male users on the Times of India website.[8] The effect was highly sensitive to the ads from this particular service, and the same effect was not reproduced in several other experiments by the same authors. The cause was found to be higher CPA (cost per conversion) for

---

[6] This includes effects which are small for a single instance but have a significant cumulative effect. As Greenwald et al. observe, statistically small effects can have substantial societal impact when they apply to many people, or if they apply repeatedly to the same person. (Anthony G. Greenwald, Mahzarin R. Banaji, and Brian A. Nosek. Statistically small effects of the Implicit Association Test can have societally large effects. Journal of Personality and Social Psychology, 108:553–561, 2015.)

[7] Note that even small reinforcement biases in systems can be magnified via positive feedback loops, since biased representations can influence human behavior which is in turn fed back into training data for those systems.

[8] Amit Datta, Michael Carl Tschantz, and Anupam Datta. Automated Experiments on Ad Privacy Settings. PETS 2015, pp. 92-112, June 2015.

female users for one campaign (notably with a higher CTR for female users) and advertiser targeting to male-only users for the other.[9]

In scope for algorithmic unfairness. In the first case, bias in user behavior trained the system to have biased targeting. In the second case, advertiser-selected criteria were related to a sensitive characteristic.

Facebook Computation of Unregulated FICO Scores

Researchers raised concerns in 2015 about Facebook computing non-regulated credit scores based on user activity.[10] One substantial issue they commented on is that credit scores are regulated by the Equal Credit Opportunity Act of 1974, which prohibits creditors from discriminating against applicants on the basis of race, religion, national origin, sex, marital status, age, or receiving public assistance; however, the non-regulated scores did not appear to have such restrictions.[11]

In scope for algorithmic unfairness. Based on the user's behavior, the system automatically computed sensitive characteristics that could adversely affect their financial opportunities.

Chumhum Suppression of Businesses in High-Crime Areas

An episode of the primetime television drama "The Good Wife" featured a tortious interference case against a fictional search engine company (Chumhum) for releasing a maps application that suppressed businesses in high crime areas.[12]

In scope for algorithmic unfairness. Content was excluded by an algorithm in a way that disproportionately affected people who owned businesses in neighborhoods correlated with the sensitive characteristics race and income.

West African Spam Filters

An external researcher poses the following: "One question is whether the design of spam filters could make certain individuals more susceptible to having their legitimate messages diverted to spam folders. For example, does being located in a hotbed of Internet fraud or spam activity, say West Africa (Nigeria or Ghana) or Eastern Europe, create a tendency for one's messages to be mislabeled as spam?"[13]

In scope for algorithmic unfairness. In this hypothetical, the system learns associations and downgrades or excludes content related to sensitive characteristics such as national origin and race.

---

[9] Giles Hogben, Alex McPhillips, Vinay Goel, and Allison Woodruff. Allegations of Algorithmic Bias: Investigation and Meta-Analysis. September 2016. go/allegations-of-algorithmic-bias

[10] Tressie McMillan Cottom. Credit Scores, Life Chances, and Algorithms. May 30, 2015.

[11] Astra Taylor and Jathan Sadowski. How Companies Turn Your Facebook Activity Into a Credit Score: Welcome to the Wild West of Data Collection Without Regulation. *The Nation*, May 27, 2015.

[12] The Good Wife, Episode "Discovery". CBS, first aired November 22, 2015.

[13] Jenna Burrell. How the machine "thinks": Understanding opacity in machine learning algorithms. Big Data & Society 3(1):1-12, 2016.

Google Photos Use of Gorillas Label

In June 2015, a web developer posted on Twitter that Google Photos had tagged an image showing him and a friend at a concert with the label "Gorillas". "Of all terms, of all the derogatory terms to use," Alciné said later, "that one came up." According to a post mortem, Google executives noticed Alciné's tweets within one hour. A Googler reached out through Twitter for permission to access the user's photos, and the issue was identified and resolved. In the short term, the Photos team stopped suggesting the "Gorillas" tile in the Explore page and stopped showing Search results for queries relevant to gorillas; in the long term, the team pledged to investigate the image annotation models that generate false positives for the gorilla label and work on the image annotation pipeline and quality evaluation processes.[14]

In scope for algorithmic unfairness. The system drew an incorrect inference that is offensive to members of a given race.

Autocomplete Results for Trayvon Martin

In 2013, Autocomplete results showed negative results for Travyon Martin (e.g., "drug dealer"), but more positive results for George Zimmerman (e.g., "hero").[15]

In scope for algorithmic unfairness. The system processed and showed offensive content from users, in a way that could be seen to reinforce existing social biases regarding race.

Google Search Results for Black Girls & Pornography

In 2013, a researcher expressed concern that search queries on Google for "black girls" historically yielded a high percentage of pornography.[16]

In scope for algorithmic unfairness. The system showed results that are offensive and reinforce existing social biases regarding race.

Google Image Search Results for Physicists

Google image search results for "physicist" show predominantly men. In reality, roughly 20% of physicists are women. First, imagine that the image search results show 1% women. Second, imagine instead that the image search results show 20% women.

In scope for algorithmic unfairness. In the first case, when the image search results show only 1% when the reality is 20%, the system is *amplifying* an existing bias in society. In the second case, when the image search results show a percentage similar to the current reality, the system is *reflecting* an

---

[14] Giles Hogben, Alex McPhillips, Vinay Goel, and Allison Woodruff. Allegations of Algorithmic Bias: Investigation and Meta-Analysis. September 2016. go/allegations-of-algorithmic-bias

[15] Safiya Umoja Noble. Trayvon, Race, Media and the Politics of Spectacle.The Black Scholar. 44(1):12-29, Spring 2014.

[16] Safiya Umoja Noble. Google Search: Hyper-visibility as a Means of Rendering Black Women and Girls Invisible. InVisible Culture: Issue 19, October 29, 2013.

existing bias in society. Both cases fall in scope for algorithmic unfairness because they reinforce a stereotype about the role of women in a scientific discipline. However, while both are in scope for algorithmic unfairness, remediation may be more likely in the former case. (As with the other cases, whether and how to remediate would fall to product policy.)

Microsoft Kinect

News media in 2010 and 2017 reported on developers and users of Microsoft's Kinect experiencing difficulties with facial recognition or motion detection for users with dark skin in low light conditions. Affected users could not use certain system functions including automatic sign in to the user's profile,[17] or use software that required motion detection to function.[18]

In scope for algorithmic unfairness. The predictive system performed poorly for users with dark skin, which is associated with the sensitive characteristic race. Users considering purchasing the Kinect or software that requires the Kinect would be unlikely to suspect that their skin color might make it difficult or impossible to use advertised features (such as automatic login using facial recognition, gestural menu interactions, motion controlled gaming). Gaming hardware and software purchases are commonly non-refundable after being opened, so a user who realized the Kinect or its software could not track them would be left with a non-functional product they could not return.

Survivalist Game on the Play Store

In January of 2016, a series of Twitter posts and a Change.org Petition raised concerns about a Survivalist game on the Play Store, claiming it gamified killing aborigines.

Not in scope for algorithmic unfairness. There were concerns that the content was offensive, but an algorithm was not involved.

Airbnb Acceptance Rates

Researchers from the Harvard Business School conducted an experiment on Airbnb and found that applications from guests with distinctively African-American names were less likely to be accepted relative to identical guests with distinctively white names. The study could not identify whether the unfairness was based on race, socioeconomic status, or some other factor.[19]

Not in scope for algorithmic unfairness. The unfairness was on the part of individual users ("sellers" on Airbnb) and there was no algorithmic component to the unfairness. While this is out of scope for algorithmic unfairness, Airbnb did however recognize that their design choices unrelated to algorithms could be facilitating unfair behavior. Accordingly, it pursued non-algorithmic remediations, such as modifying its user interface design and its end user policy.[20]

---

[17] Brendan Sinclair. Kinect has problems recognizing dark-skinned users? Gamespot, 2010.

[18] Andy Trowers. How We Accidentally Made a Racist Videogame. Kotaku, 2017.

[19] Benjamin Edelman, Michael Luca, and Dan Svirsky. Racial Discrimination in the Sharing Economy: Evidence from a Field Experiment. Harvard Business School NOM Unit Working Paper (16-069), 2015.

[20] Laura W. Murphy. Airbnb's Work to Fight Discrimination and Build Inclusion: A Report Submitted to Airbnb. September 2016.

FAQ

Is personalization the same as algorithmic unfairness?
Generally, no. Personalization is algorithmic behavior that presents different results to different users. Personalization is very often beneficial to users and is not necessarily unfair. In fact, personalization may sometimes remediate unfairness by identifying an individual's interests more precisely than would the use of a sensitive characteristic.[21] However, in some cases personalization may be strongly associated with a sensitive characteristic in a way that causes significant harm to a user or users, in which case it would be algorithmic unfairness.

Can any characteristic be sensitive?
No. For example, "people who like yellow" or "people with pets" are unlikely to be sensitive.[22]

Which characteristics are sensitive?
Sensitive characteristics may be determined by legal considerations or by more broad principles. Legally determined characteristics may vary by jurisdiction, sector, or other factors. Sensitive characteristics determined by more broad principles are especially likely to include characteristics that are associated with less privileged or marginalized populations (particularly when such characteristics are immutable), may be socially undesirable, or may be associated with civil liberties.

Specific examples may include race/ethnic origin, gender identity, sexual orientation, religion, political party, disability, age, nationality, veteran status, socioeconomic status (including caste and homelessness), and immigrant status (including refugee and asylum seeking status). Further, there may be emergent sensitive characteristics for which we do not yet have a name, or which we have not yet anticipated.[23]

Are proxies for sensitive characteristics covered?
Yes. Proxies are characteristics that are highly correlated with a sensitive characteristic. Actions based on close proxies for sensitive characteristics are in scope for further investigation. For example, if interest in hip-hop music is highly correlated with being Black, targeting users who like hip hop music may result in algorithmic unfairness based on race, regardless of intent.

If a system's behavior is caused by societal bias, can it still be algorithmic unfairness?
Yes. Societal bias that is reflected in algorithmic behavior is a central issue in the external public, media, and regulatory concerns about algorithmic unfairness.[24,25,26,27] Additionally, one often doesn't

---

[21] James C. Cooper. Separation and Pooling. George Mason Law & Economics Research Paper No. 15-32, March 2016.
[22] http://civilrights.findlaw.com/civil-rights-overview/what-is-discrimination.html
[23] danah boyd, Karen Levy & Alice Marwick, The Networked Nature of Algorithmic Discrimination. In Seeta Peña Gangadharan, Virginia Eubanks, and Solon Barocas (eds), Data and Discrimination: Collected Essays. Washington, D.C.: Open Technology Institute, New America Foundation, pp. 53-57, 2014.
[24] Tarleton Gillespie. The Relevance of Algorithms. In Media Technologies: Essays on Communication, Materiality, and Society, edited by T. Gillespie, P. Boczkowski, and K. Foot. MIT Press, Cambridge, MA, 2012.
[25] Solon Barocas and Andrew D. Selbst. Big Data's Disparate Impact. 104 California Law Review 671 (2016).
[26] Executive Office of the President. Big Data: Seizing Opportunities, Preserving Values. May 2014.

know the root cause of algorithmic unfairness without an investigation, so any real-world process (e.g., incident response) is best served by encompassing a wide range of potential causes. Accordingly, this definition covers a wide range of sources, from machine learning classification errors to societal bias. However, the nature of the root cause may affect product policy's position on whether or how a given algorithmic behavior should be addressed.

If a representation is factually accurate, can it still be algorithmic unfairness?
Yes. For example, imagine that a Google image query for "CEOs" shows predominantly men. Even if it were a factually accurate representation of the world, it would be algorithmic unfairness because it would reinforce a stereotype about the role of women in leadership positions. However, factual accuracy may affect product policy's position on whether or how it should be addressed. In some cases, it may be appropriate to take no action if the system accurately affects current reality, while in other cases it may be desirable to consider how we might help society reach a more fair and equitable state, via either product intervention or broader corporate social responsibility efforts.

If a system's behavior is not intended, can it still be algorithmic unfairness?
Yes. If the behavior is unfair, it meets the definition regardless of the root cause.

If unfairness is executed by an algorithm but is the result of a human decision, can it still be algorithmic unfairness?
Yes. For example, if a human chooses unfair keywords for ads and those choices result in unfair algorithmic choices of what ads to show to a user or users, that would fall within the scope of algorithmic unfairness.

Does this definition include the use of data by multiple parties?
Yes. For instance, decisions which may result in data being provided to third parties via APIs, being sold to third parties, or acquired via federated identity should consider the possibility that that data could be used to power unfair algorithmic decision-making.

What is the relationship between this definition of algorithmic unfairness and the fairness measure in Hardt et al. (2016)?
The fairness measure in Hardt et al. (2016)[28] is a statistical guarantee that, in general, members of one category are classified with the same accuracy as members of another category. For example, if Black people who apply for credit cards are classified with lower accuracy than white people (e.g., Black people who would actually repay their loans are wrongly classified as bad credit risks), the fairness measure in Hardt et al. (2016) would find that to be a problem. However, if most people who happen to have a particular sensitive characteristic are correctly classified as being very likely to default on their credit cards (regardless of whether that classification is done based on that sensitive characteristic or based on some other associated variable), the Hardt et al. (2016) measure would not detect it as an issue. Therefore, the fairness measure in Hardt et al. (2016) statistically tests for a certain type of machine learning classification error that would constitute a specific type of algorithmic unfairness, encompassing some but not all of the cases covered by the current definition.

---

[27] ACM US Public Policy Council. Statement on Algorithmic Transparency and Accountability. January 12, 2017.
[28] Moritz Hardt, Eric Price, and Nathan Srebro. Equality of Opportunity in Supervised Learning. arXiv.org, October 2016.

Acknowledgments

More on misinformation and our role within it.

---------- Forwarded message ---------

[redacted]

Fake News-letter

BY gTech Search + Content
POC: valstreiter

NOVEMBER 27, 2017 | INTERNAL ONLY

SOURCE QUALITY

Revamping News Corpus

Goal: Establish "single point of truth" for definition of "news" across Google products. Mitigate risk of low-quality sources and misinformation in Google News corpus.
Status: Define new Google News corpus utilizing existing infrastructure and tools while integrating different quality tiers and labels, algo-human content review, new cross-product review/exclusion pipeline and updated inclusion UI.

CONTACTS [redacted]
MORE INFO: Proposal, Updated rater template (WIP) / Stress Test publisher list
NEXT STEPS: Research and gather info about possible signals to leverage for product teams to make source quality decisions. Finalize and test new rater inclusion guidelines with standards to combat misinformation.

Trust Project & Nutrition Labels

Goal: Develop transparency standards that help people easily assess quality and credibility of journalism. Work across News, Nutrition Labels, and Search Console / 3P structured data teams to incorporate Trust Project data into future product plans.
Status: Launched eight trust indicators with a dozen publishers going live with schema implementations. Launched publisher Knowledge Panels in Search.

CONTACTS [redacted]
MORE INFO: Publisher KP launch, Announcement, The Trust Project
NEXT STEPS: Complete PRD Addendum and SD Playbook for markup support in Search Console.

Reviewing Rater Quality

Goal: Evaluate general performance of Ewok rater evaluations in select countries for News corpus inclusion to uncover potential red flags or unusual patterns.

Status: Analyze low-quality and high-quality rater comments, examine inconsistencies and rater inclusion history.

CONTACTS: [redacted]
MORE INFO: A/C Privileged and Confidential

PRODUCT OPERATIONS

Project Purple Rain: Crisis Response & Escalation

Goal: Establish and streamline news escalation processes to detect and handle misinformation across products during crises. Install 24/7 team of trained analysts ready to make policy calls and take actions across news surfaces including News, News 360 and Feed.
Status: SOS Alerts, Crisis Response, HotEvent, and T&S Incident Management teams are collaborating to identify a narrow set of queries that would be used to manually trigger flight-to-quality in Search. T&S Incident Management team is currently looking to expand and share resources with teams that currently handle Suggest and WebAnswers escalations.

CONTACTS: [redacted]
MORE INFO: Slide deck, Meeting notes, Mailing list

Policy Development & Enforcement

Goal: Develop and proactively enforce misformation policy that scales across News products. Explore cross-product toolset for removal enforcement.
Status: Drafted policy for Coordinated Inauthentic Cross-Border Information Operations (CICBIO). Reviewed proposed policy with Search & News leadership.

CONTACTS: [redacted]
MORE INFO: T&S proposal to expand policy enforcement
NEXT STEPS: Set up workflows and tools with TAG team and other stakeholders. jeffreyc@ and @jacobhelberg@ to scope "badness baseline rating" and other tools for top of P0 escalation funnel.

ESCALATIONS

Italian sites spread misinformation: PR escalation following Buzzfeed article. 3 sites rejected following ewok and manual evaluations.

Hoax science stories in News: Junk science articles flagged by internal Googlers and written up in blog. Status pending: 5 sites submitted for ewok re-evaluation, 3 rejected.

Dylann Roof sentencing resurfaces in News: Danny Sullivan's sleuthing uncovered cause as publisher error that spawned copycats. (Resolved on Twitter)

CA shooting rampage: No major issues across products. PR post mortem with open questions and follow-up here.

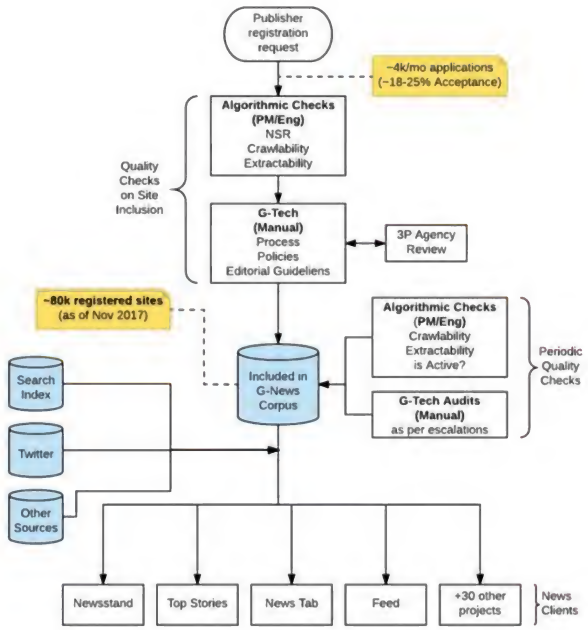
TX church shooting: We surfaced some misleading tweets in Search and some low-quality videos in YouTube which resulted in a negative press cycle. Danny Sullivan responded with a series of tweets. (PR summary)

Expanding Collaboration for News Quality

Fatih Ozkosemen, Trust & Safety Search

History

News Ecosystem



Prev. Work on Web Spam

Feeding WebSpam Manual Actions to News Corpus

- Design Doc
- Corpus removal
  - Removal, demotion penalties
- Temp. corpus suspension
  - Hacked

Potential

Global Partnerships

Google efforts to address Fake News

What Google is doing to address the problem
go/FakeNews

Gustavo Fuentes-Rivera, Pauline Peyronnet, Mark Lyall

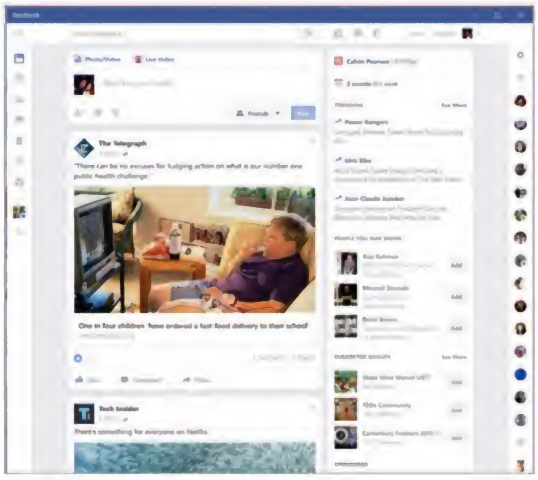
Dec. 2016

Users have access to a large variety of news sources

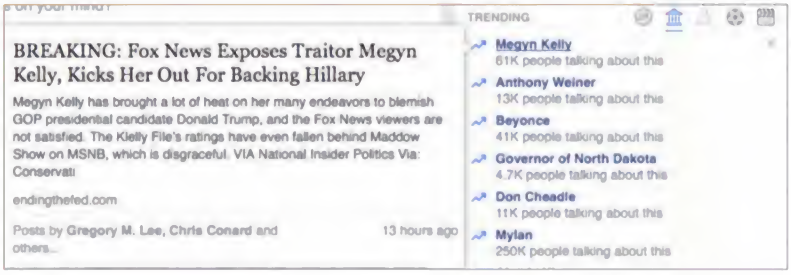
The web offers a large variety of options.

From reputable, traditional news sites to newer, relatively unknown news outlets.

Many users access news sites via FB News Feed

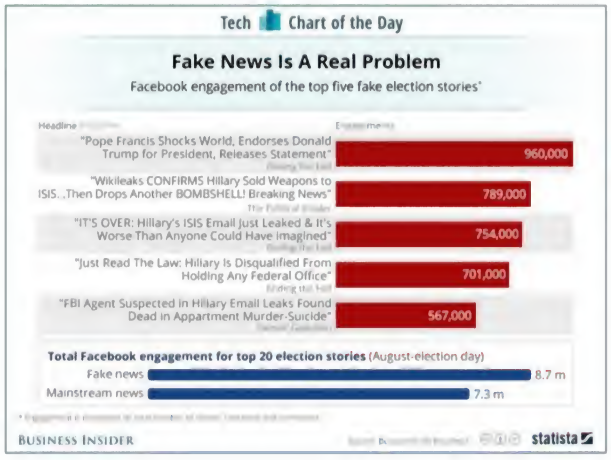


but some bad actors came into play



In the context of the US elections, a number of bad actors used the Facebook News Feed to promote fake news sites

And the issue gained notoriety



We discovered that some of these sites were using AdSense to monetize their traffic

and we took quick action, updating our AdSense policies to address 'Misrepresentative content'

With this update we disallow sites that mislead users

> **Approved Public Statement:** *We've updated our publisher policies and will start prohibiting Google ads from being placed on misrepresentative content, just as we disallow misrepresentation in our ads policies. Moving forward, we will restrict ad serving on pages that misrepresent, misstate, or conceal information about the publisher, the publisher's content, or the primary purpose of the web property.*

This update solved the problem of fake news sites monetizing with AdSense.
We're working on a similar update for AdX to cover all our bases.
Additionally there is a working group established to assess whether changes are needed for our platform products (DBM, DFP)

But then another problem surfaced: Ads

We received escalations from publishers about ads promoting misleading information/sites.

How are we addressing this?

(i) Reviewing suspect ads and taking action in cases where there is a violation of our ads policies.

(ii) Assessing what improvements can be done in automation support and/or review procedures

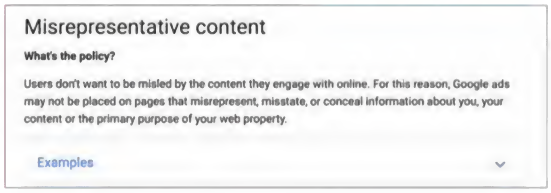
(iii) Preparing guidance for publishers on what steps can they take on AdX to block ads that they consider undesirable

Resources for external communications

"Misrepresentative Content" update on AdSense Policy - Comms Doc

> **Approved Public Statement:** *We've updated our publisher policies and will start prohibiting Google ads from being placed on misrepresentative content, just as we disallow misrepresentation in our ads policies. Moving forward, we will restrict ad serving on pages that misrepresent, misstate, or conceal information about the publisher, the publisher's content, or the primary purpose of the web property.*

AdSense Program Policies - Prohibited content



Additional resources (internal)

Internal FAQ for AdSense Misrepresentative Content Policy

Summary

Two key areas of concern:

1. **Monetization**
   What is Google doing:
   We have updated our AdSense policies to restrict ads on sites that misrepresent, misstate, or conceal information about the publisher.
   To cover all bases we will make a similar update to AdX policies.

2. **Ads**
   What is Google doing:
   > We are reviewing suspect ads and taking action in cases where there is a violation of our ads policies.
   > We are assessing what improvements can be done in automation support and/or review procedures
   > We are preparing guidance for publishers on what steps can they take on AdX to block ads they consider undesirable

Beyond Web Spam

Increasing policy coverage over

- Abuse

Other spam types, malware, deception etc.

- Fringe/controversial

Factually incorrect, fake, irrelevant, non-news etc.

- Sensitive

Hate, geo-politically sensitive, diversity & bias etc.

- Inappropriate

Sexually explicit, graphic violence and vulgarity etc.

T&S can help

- Badness signal discovery & data sharing for policy enforcement
- Multi-lang market analyst support across all top languages globally
- Leveraging TVC resources where automation is not possible
- Aligning with policy (GPP is part of T&S)
- T&S Eng resources

Proposed AIs

- setting up a metric for corpus quality

News corpus badness rate

- broader policy enforcement and abuse detection

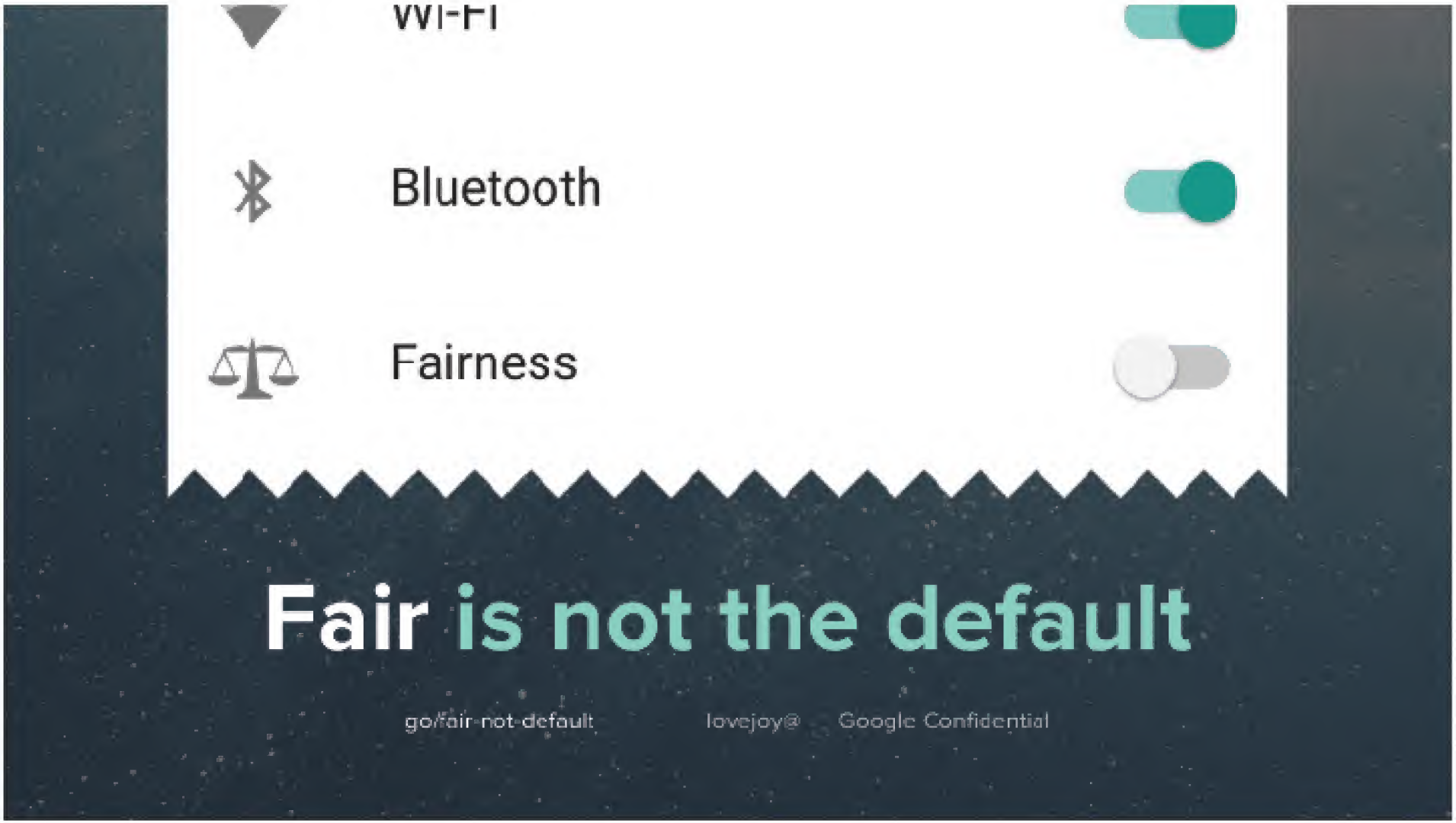
Beyond webspam

- better detecting low quality

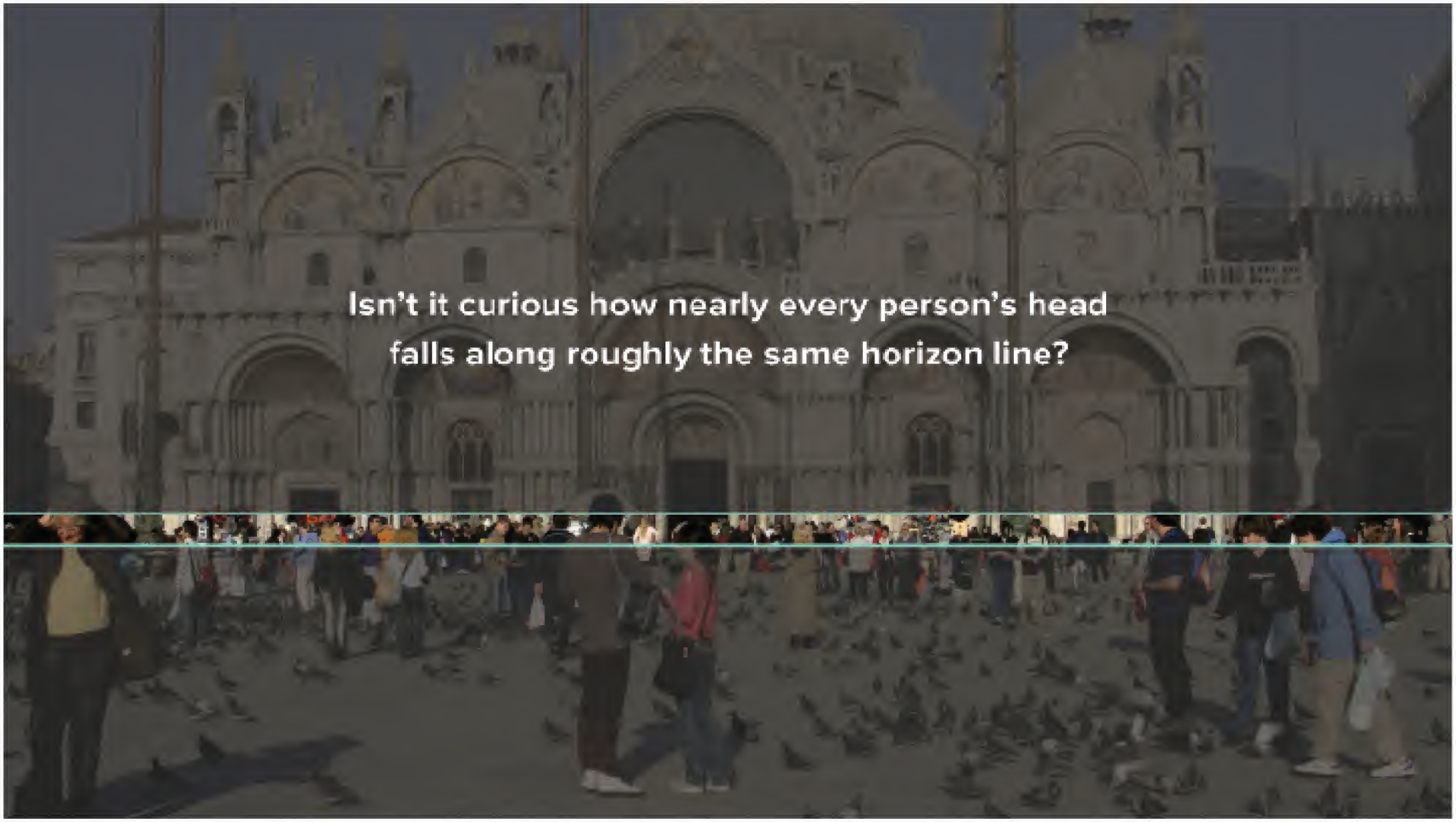
lq sources + lq content in hq sources

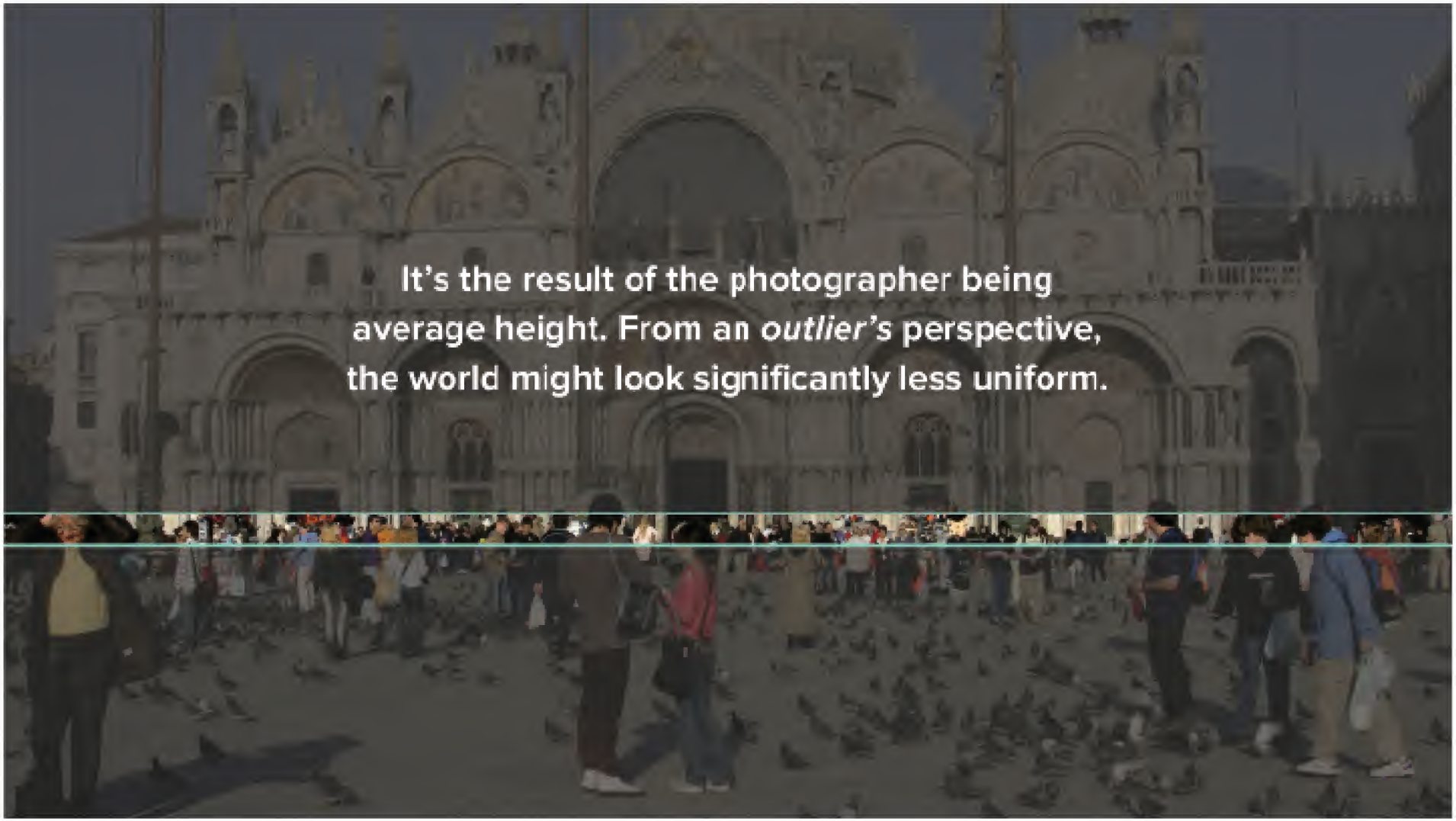
- clean & regularly sanitized news corpus

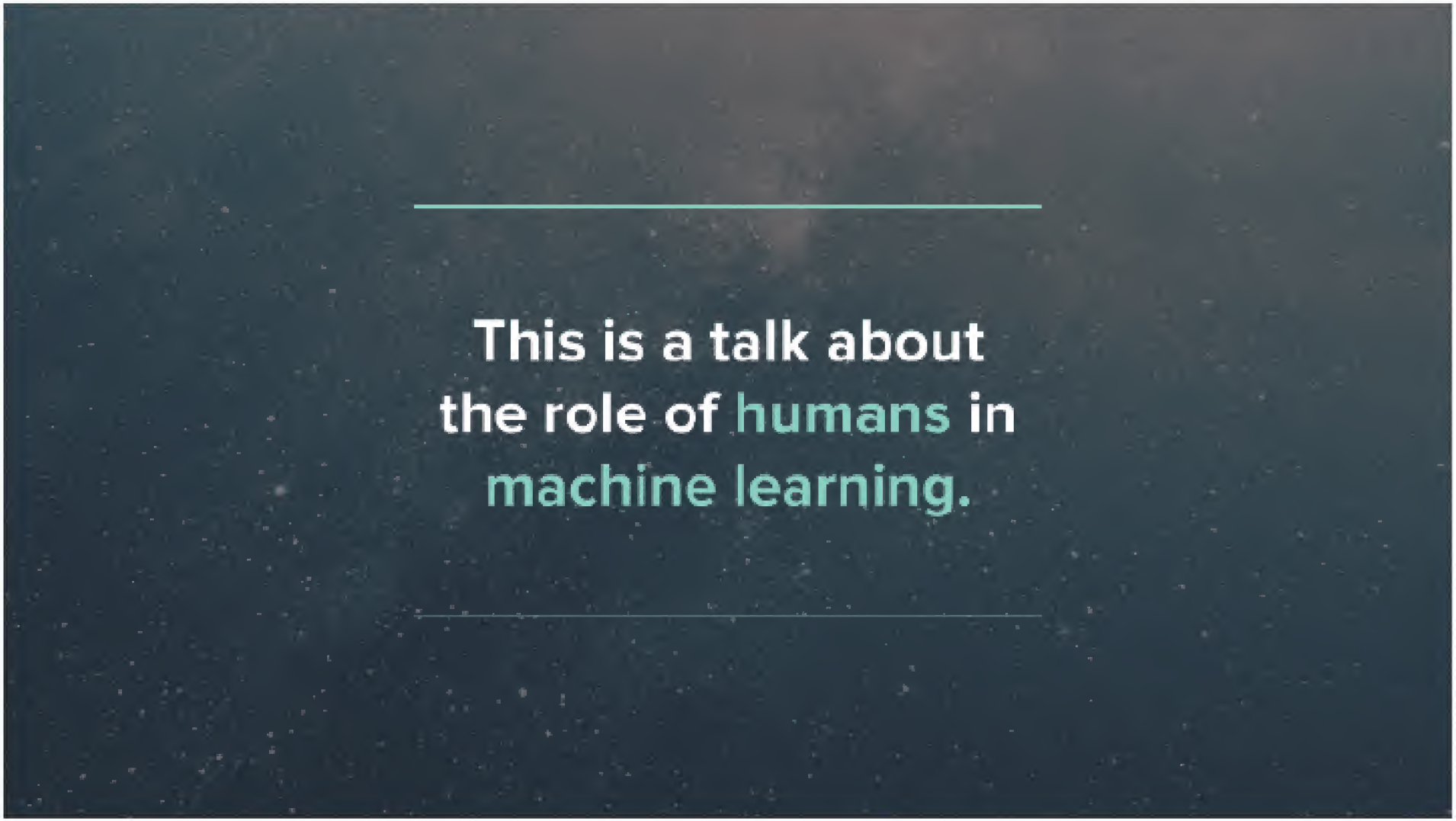
both for site registration + ongoing


Wi-Fi
Bluetooth
Fairness
Fair is not the default
go/fair-not-default
lovejoy@
Google Confidential


San Marco Square
Venice


Isn't it curious how nearly every person's head
falls along roughly the same horizon line?

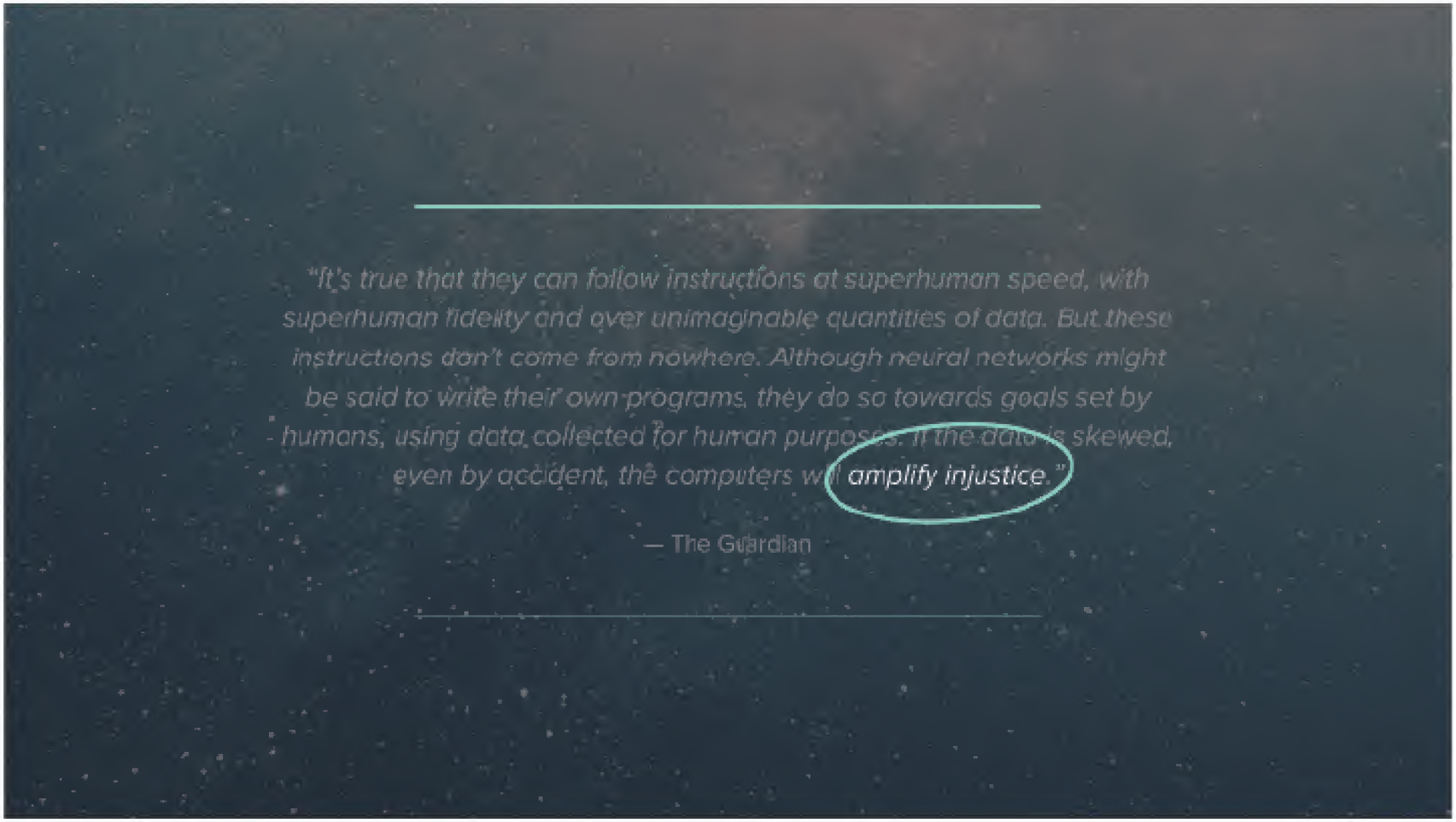

It's the result of the photographer being
average height. From an *outlier's* perspective,
the world might look significantly less uniform.

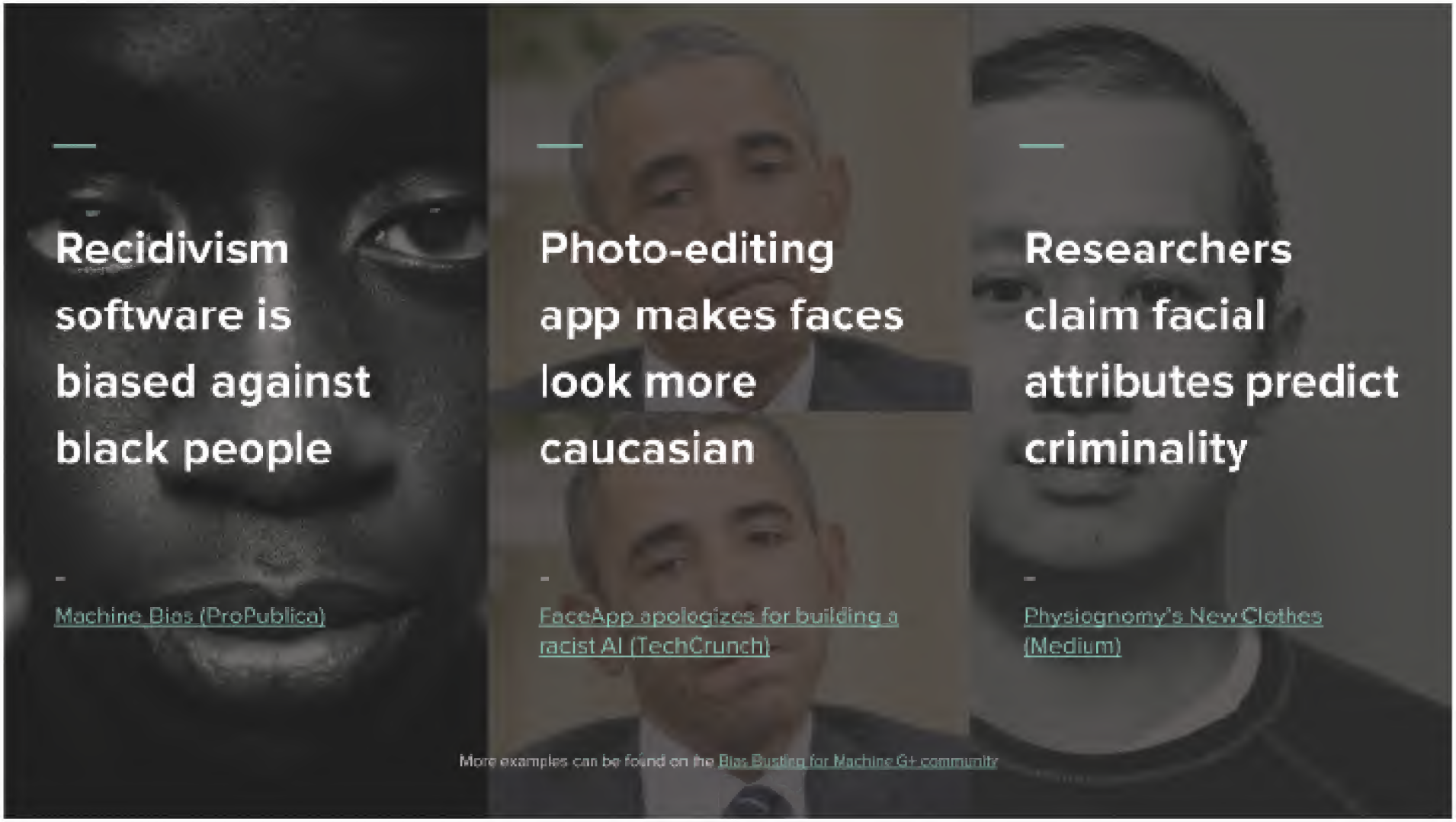

This is a talk about
the role of humans in
machine learning.

But really it's a talk about the role of humans in decision making.

"It's true that they can follow instructions at superhuman speed, with superhuman fidelity and over unimaginable quantities of data. But these instructions don't come from nowhere. Although neural networks might be said to write their own programs, they do so towards goals set by humans, using data collected for human purposes. If the data is skewed, even by accident, the computers will amplify injustice."

— The Guardian

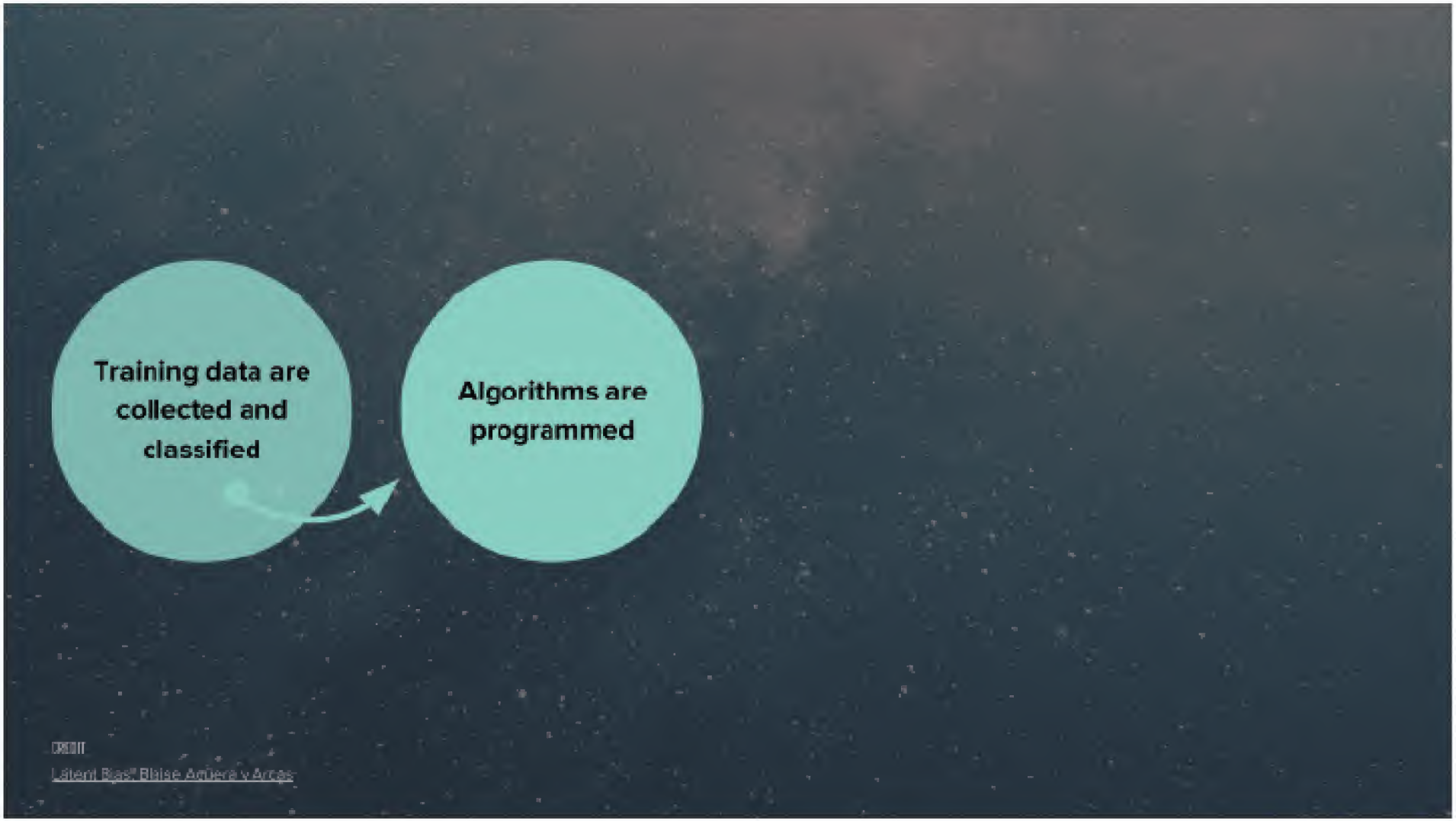

"It's true that they can follow instructions at superhuman speed, with superhuman fidelity and over unimaginable quantities of data. But these instructions don't come from nowhere. Although neural networks might be said to write their own programs, they do so towards goals set by humans, using data collected for human purposes. If the data is skewed, even by accident, the computers will amplify injustice."
— The Guardian


Recidivism software is biased against black people
Machine Bias (ProPublica)
Photo-editing app makes faces look more caucasian
FaceApp apologizes for building a racist AI (TechCrunch)
Researchers claim facial attributes predict criminality
Physiognomy's New Clothes (Medium)
More examples can be found on the Bias Busting for Machine G+ community

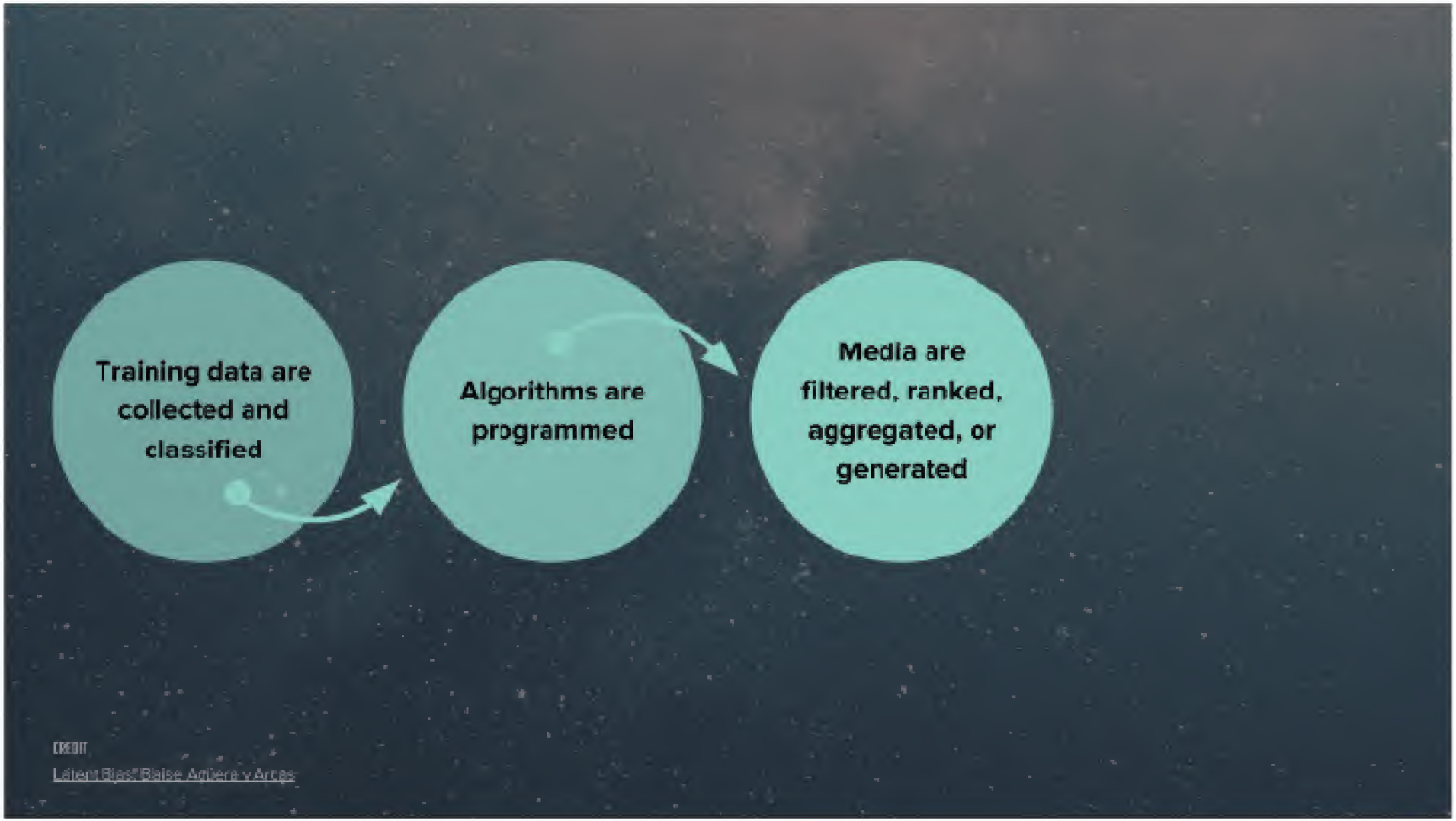
When we presume that human judgment can—or should—be removed from the loop, the result is an unconscious bias network effect.

And we (Googlers) are just as susceptible to this effect as our users.


Training data are collected and classified
CREDIT
Lakshmi Ram, Blaise Agüera y Arcas

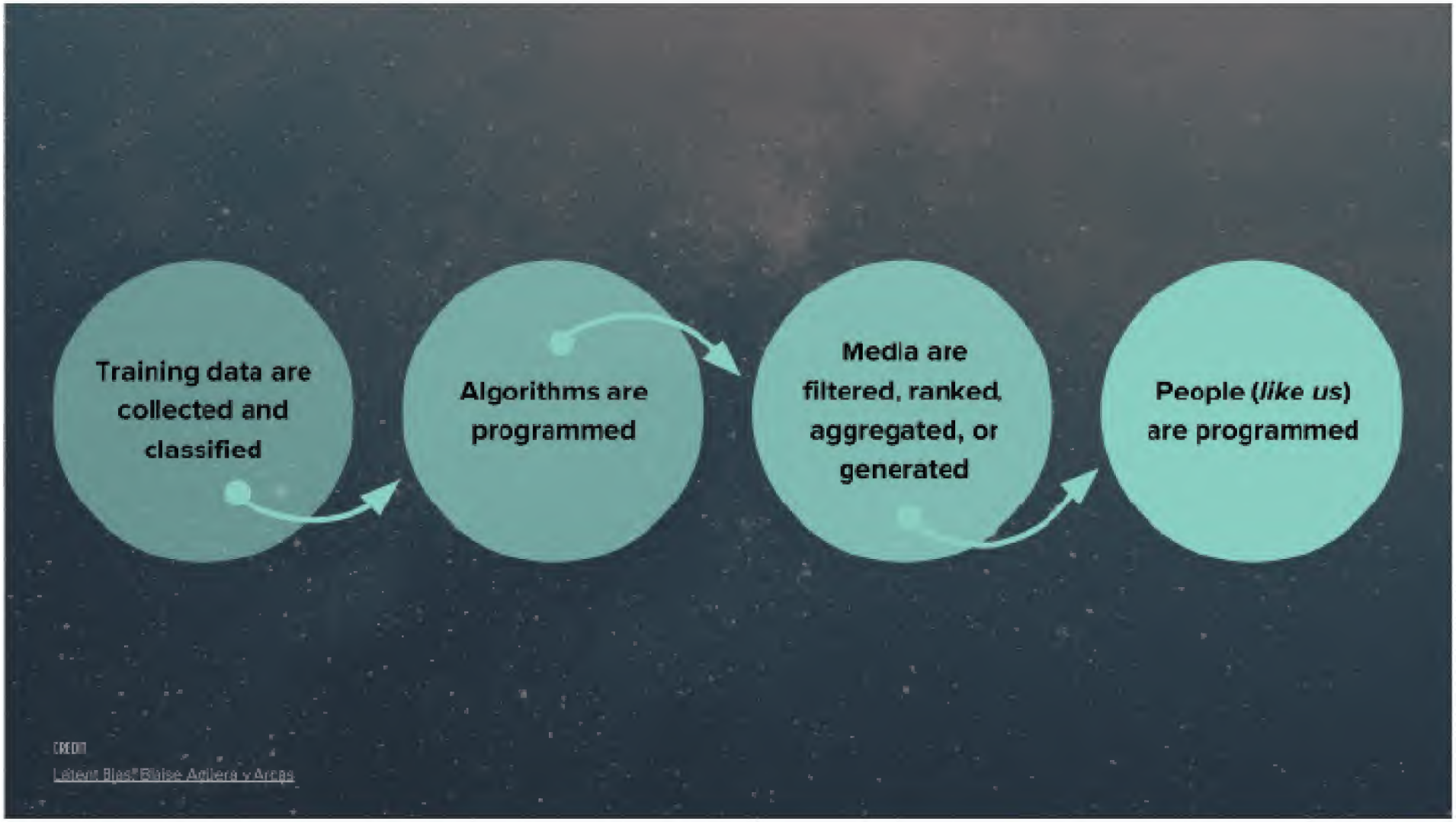


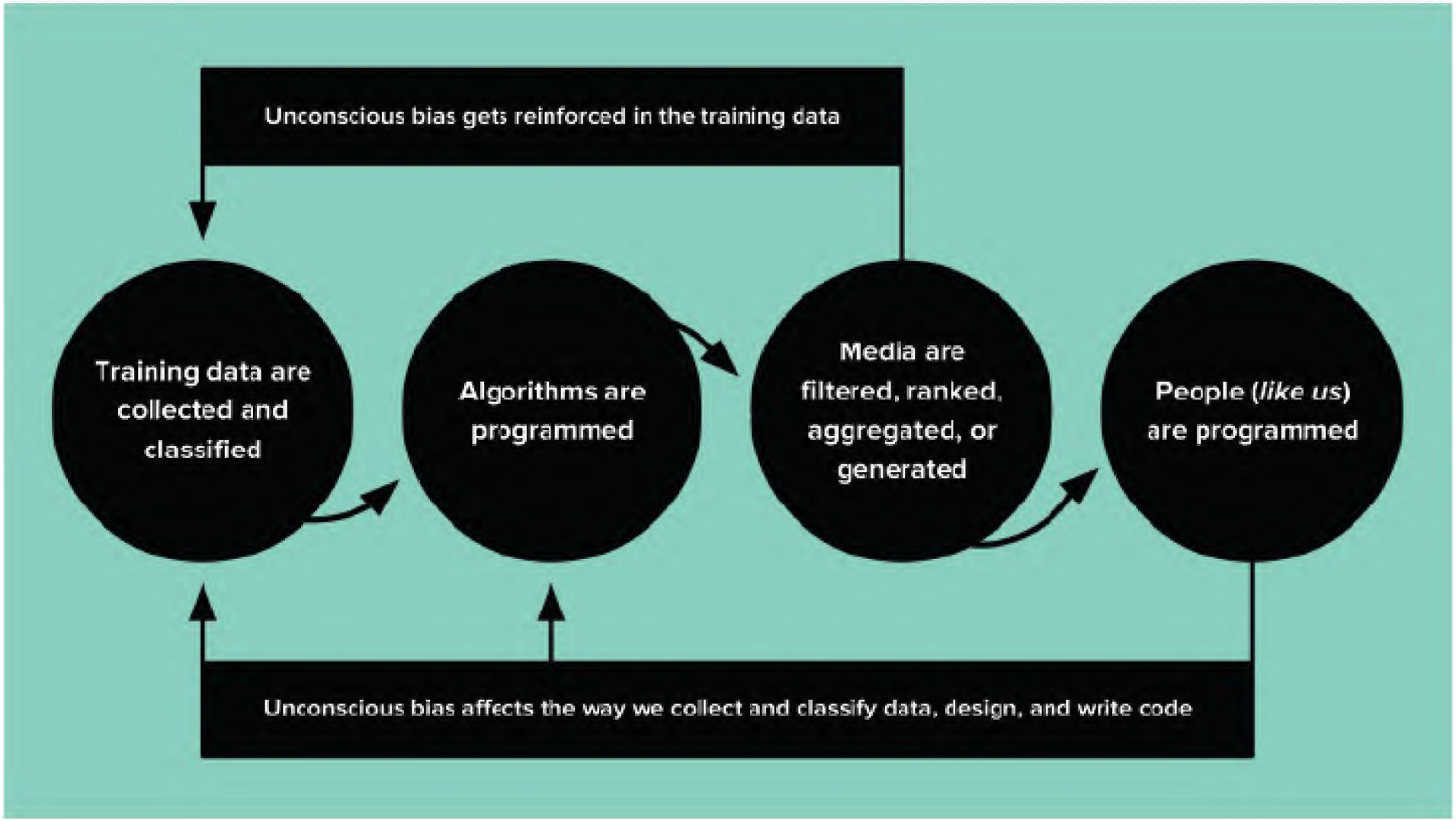
Latent Bias" Blaise Aguera y Arcas



CREDIT

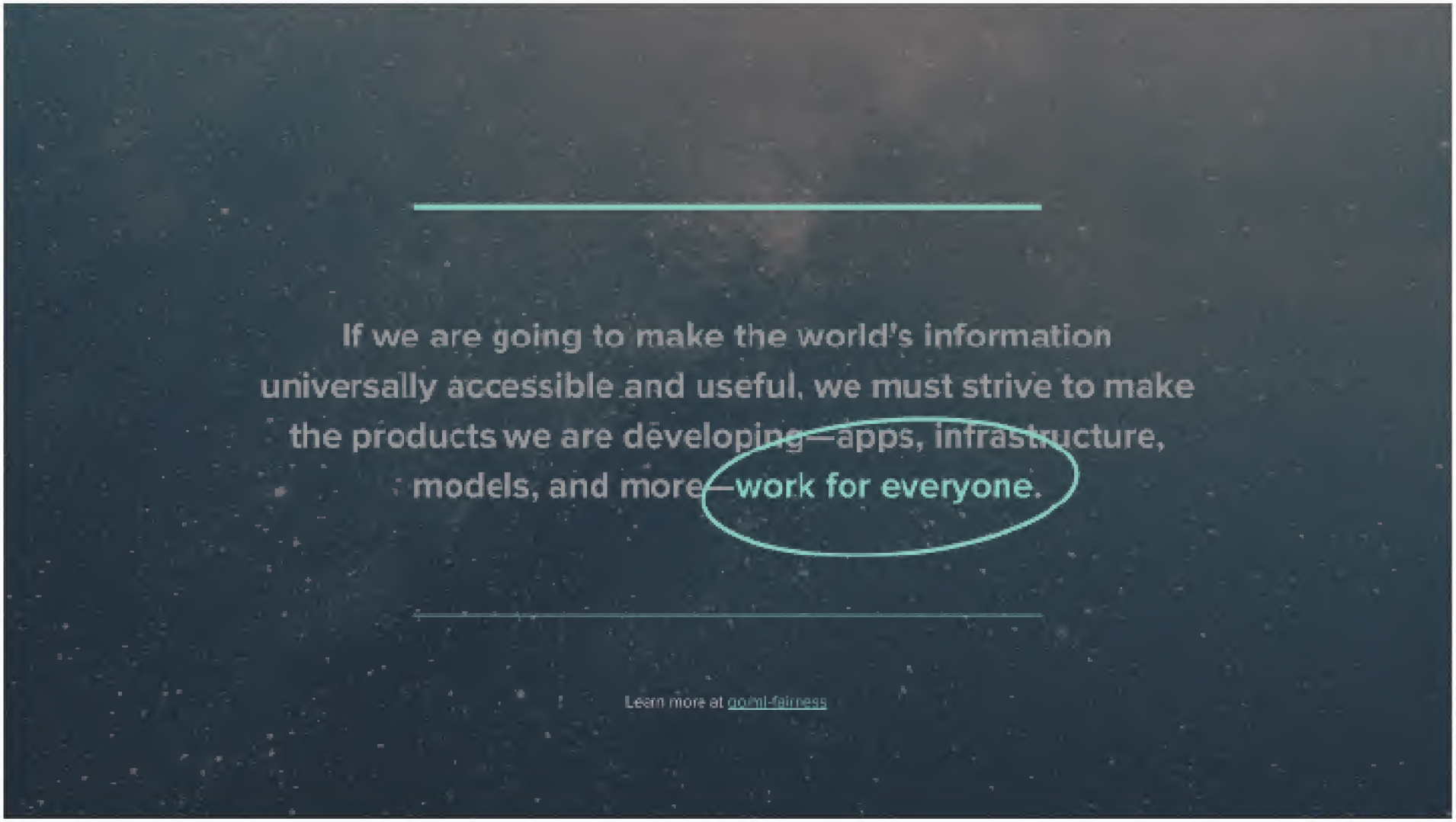
Latent Bias" Blaise Agüera y Arcas

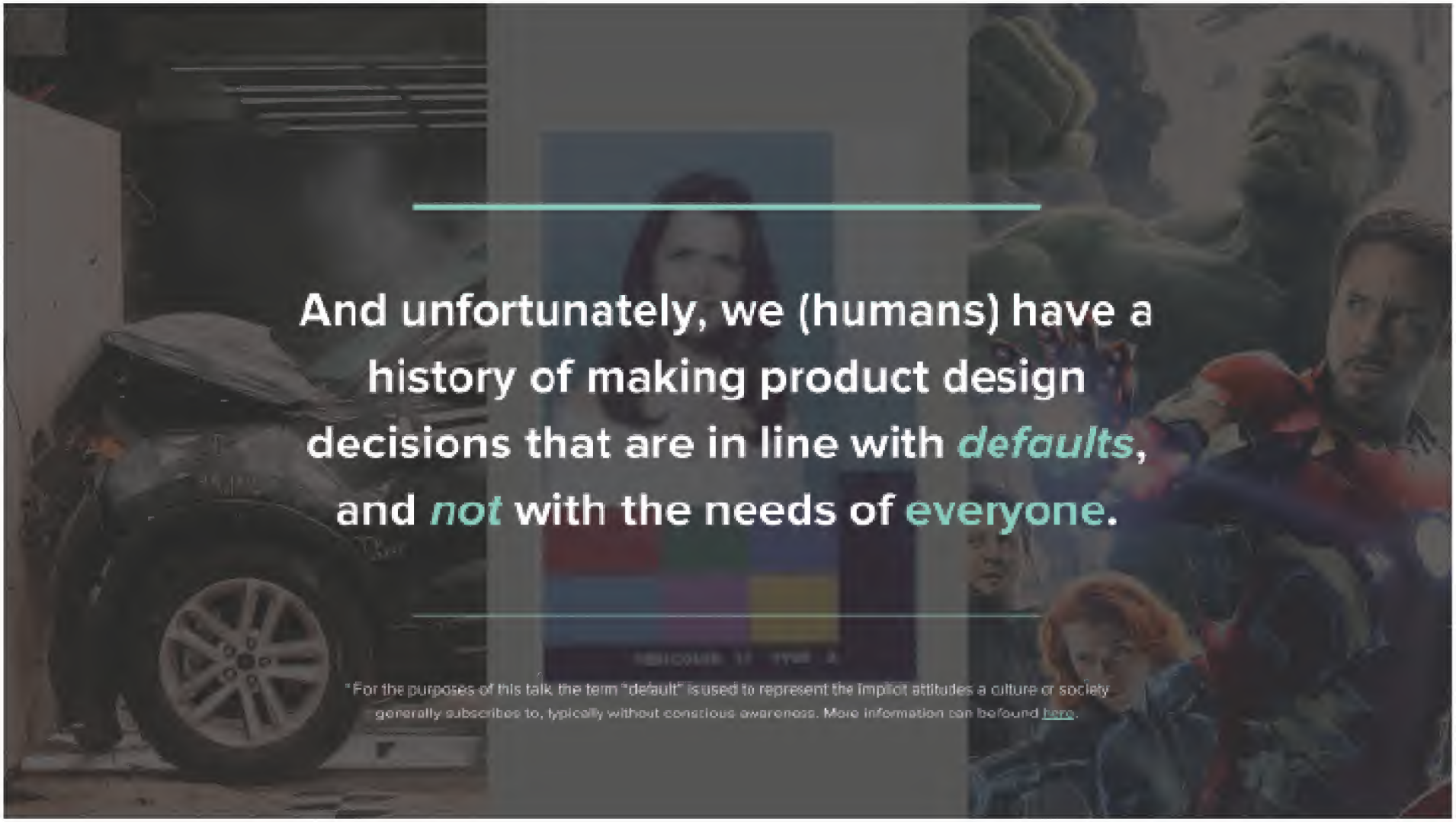

Training data are collected and classified
Algorithms are programmed
Media are filtered, ranked, aggregated, or generated
People (*like us*) are programmed
CREDIT
Latent Bias, Blaise Agüera y Arcas

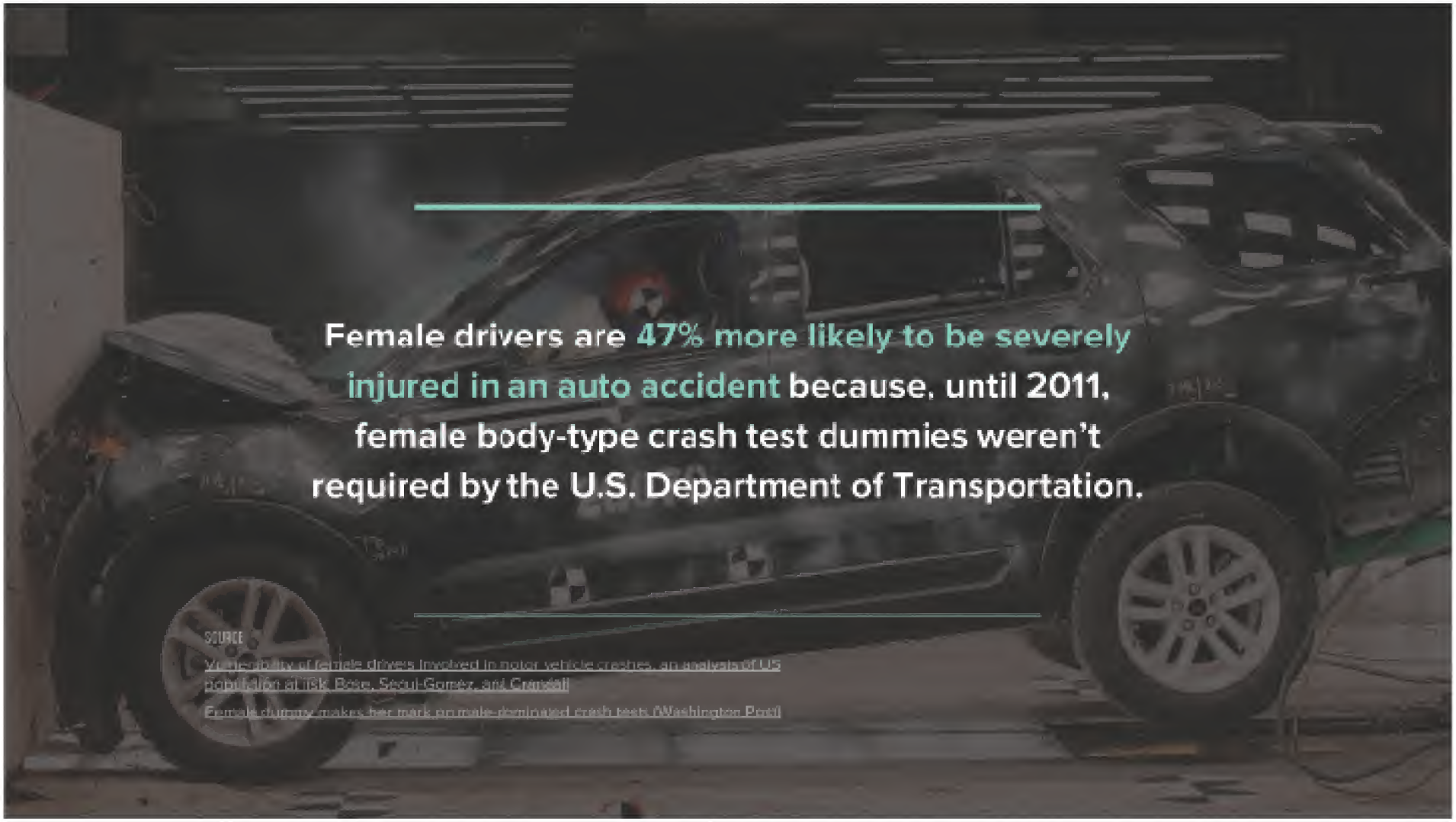

Unconscious bias gets reinforced in the training data
Training data are collected and classified
Algorithms are programmed
Media are filtered, ranked, aggregated, or generated
People (*like us*) are programmed
Unconscious bias affects the way we collect and classify data, design, and write code

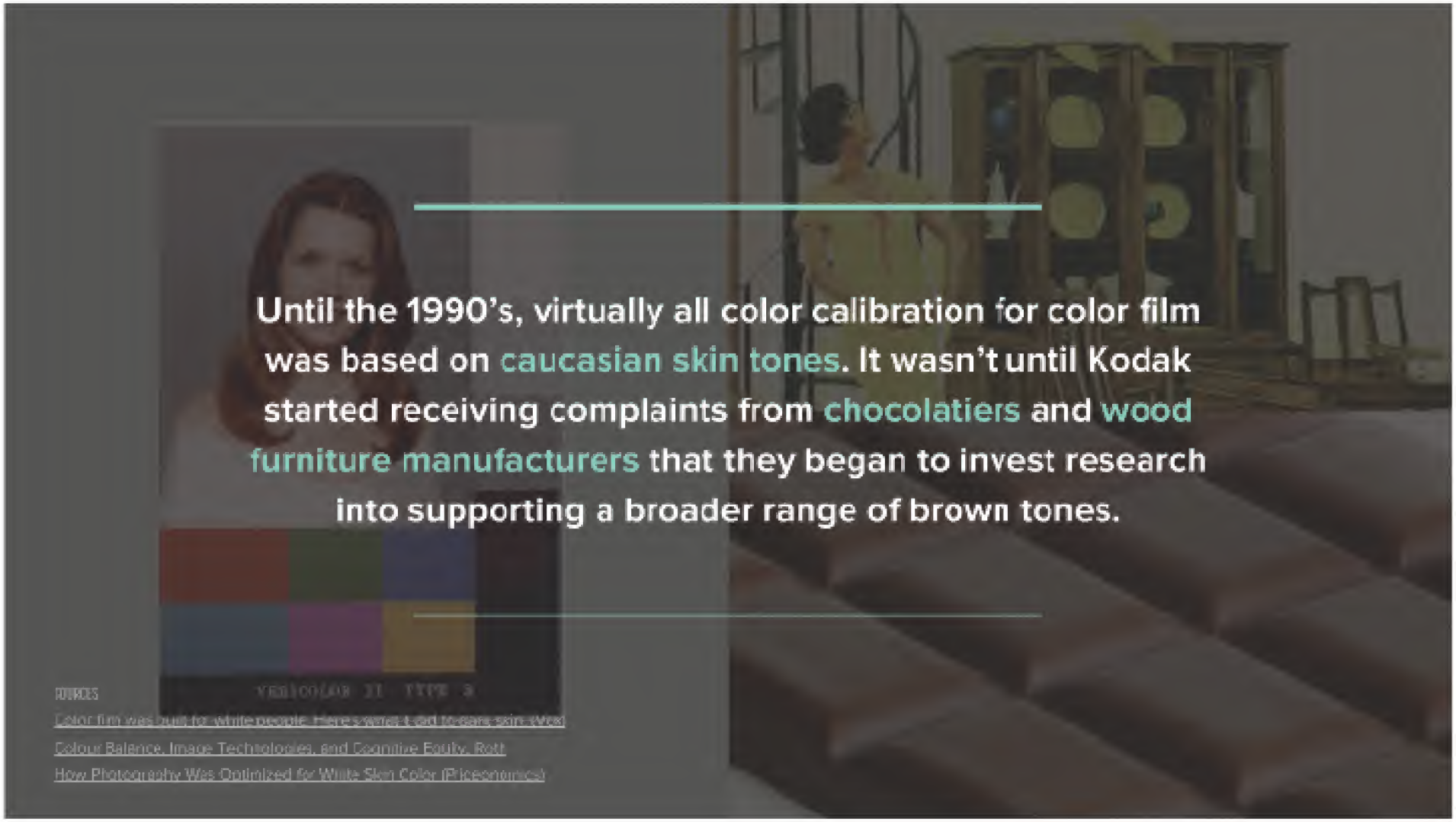
If we are going to make the world's information universally accessible and useful, we must strive to make the products we are developing—apps, infrastructure, models, and more—work for everyone.

Learn more at goo.gl/ml-fairness


If we are going to make the world's information universally accessible and useful, we must strive to make the products we are developing—apps, infrastructure, models, and more—work for everyone.
Learn more at goo.gl/ml-fairness


And unfortunately, we (humans) have a history of making product design decisions that are in line with *defaults*, and *not* with the needs of everyone.
*For the purposes of this talk the term "default" is used to represent the implicit attitudes a culture or society generally subscribes to, typically without conscious awareness. More information can be found here.


Female drivers are 47% more likely to be severely injured in an auto accident because, until 2011, female body-type crash test dummies weren't required by the U.S. Department of Transportation.
SOURCE
Vulnerability of female drivers involved in motor vehicle crashes: an analysis of US population at risk, Rose, Seoul-Gomez, and Cranwell
Female dummy makes her mark on male-dominated crash tests (Washington Post)

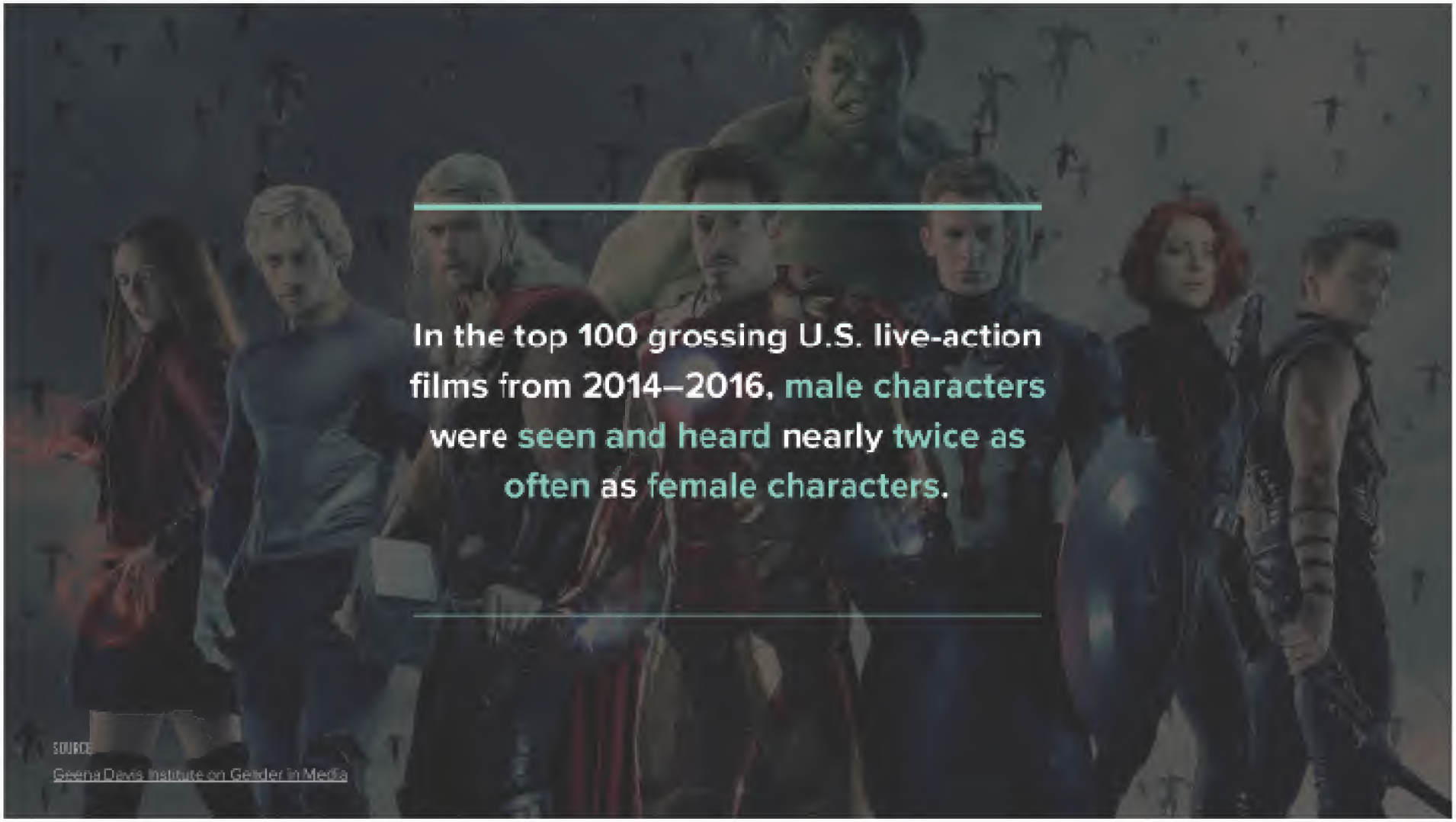


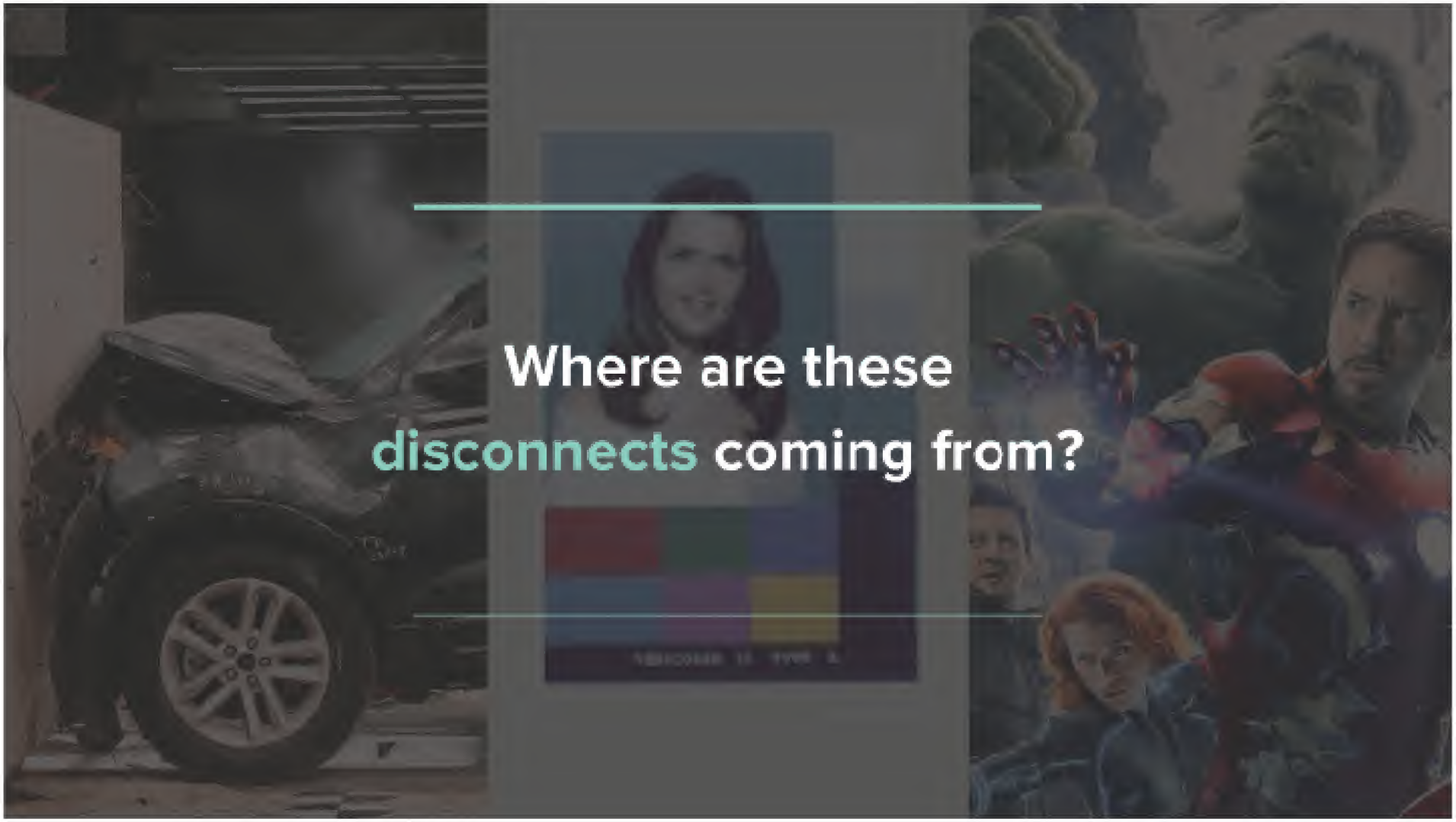
Until the 1990's, virtually all color calibration for color film was based on caucasian skin tones. It wasn't until Kodak started receiving complaints from chocolatiers and wood furniture manufacturers that they began to invest research into supporting a broader range of brown tones.

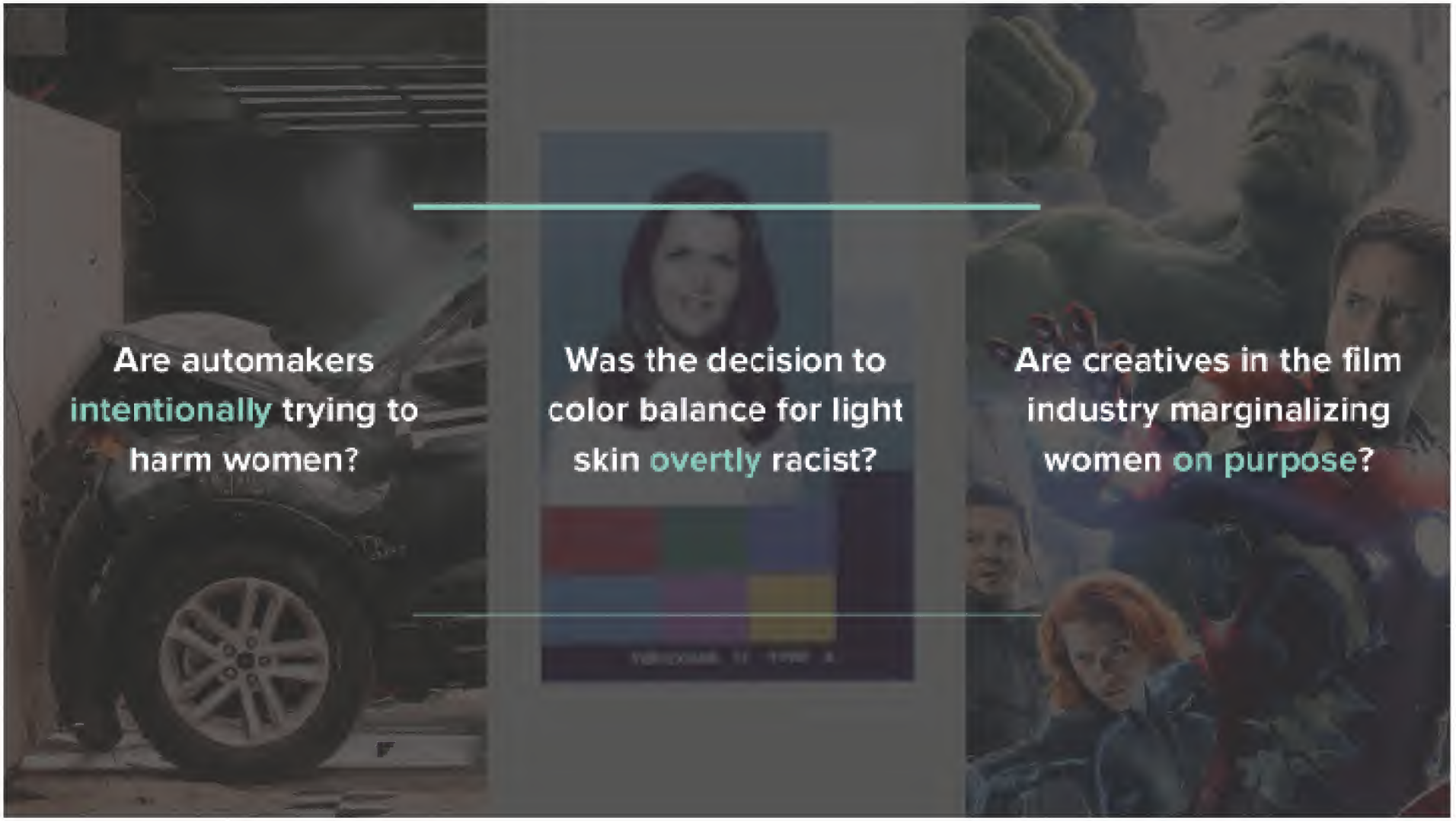
Color film was built for white people. Here's what it did to dark skin (Vox)

Colour Balance, Image Technologies, and Cognitive Equity, Roth

How Photography Was Optimized for White Skin Color (Priceonomics)

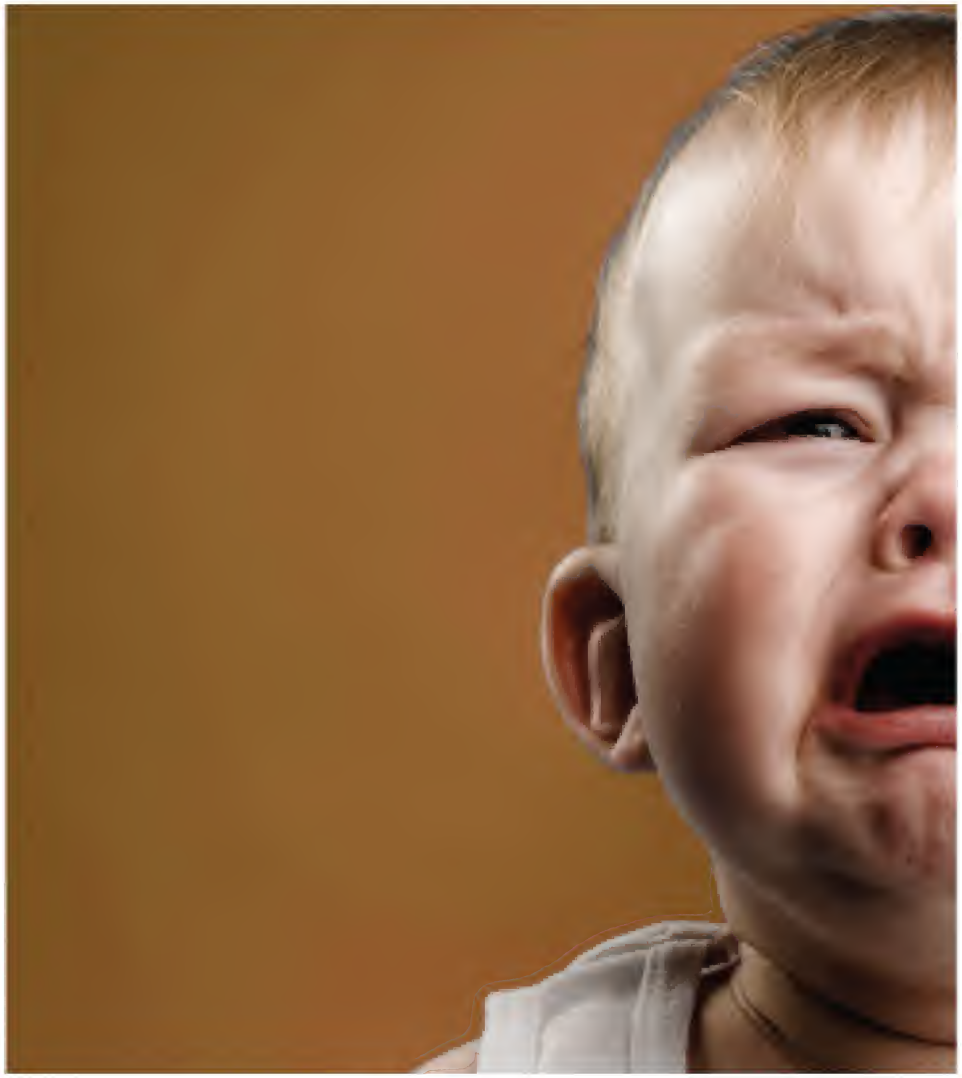

In the top 100 grossing U.S. live-action films from 2014–2016, male characters were seen and heard nearly twice as often as female characters.
SOURCE
Geena Davis Institute on Gender in Media


Where are these
disconnects coming from?

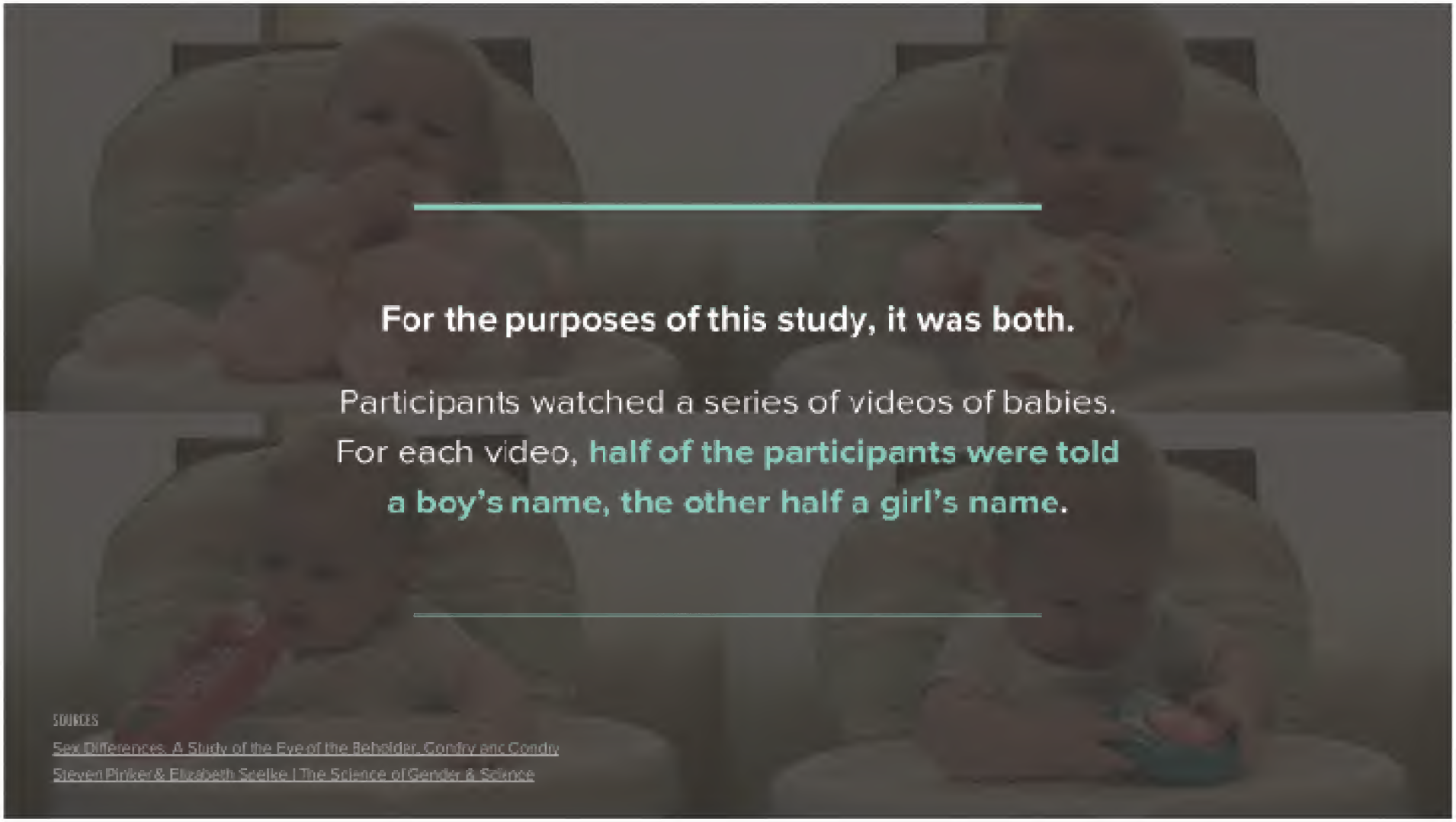

Are automakers intentionally trying to harm women?
Was the decision to color balance for light skin overtly racist?
Are creatives in the film industry marginalizing women on purpose?

These are often the behaviors of rational actors making what *seemed* like 'obvious' choices, *without* malice or ill-will.

Take for example the case of “David”



Or was it “Diana”?



For the purposes of this study, it was both.

Participants watched a series of videos of babies. For each video, **half of the participants were told a boy's name, the other half a girl's name.**

SOURCES

Sex Differences: A Study of the Eye of the Beholder, Condry and Condry

Steven Pinker & Elizabeth Spelke | The Science of Gender & Science

When the babies did something unambiguous, reports were not affected by the perceived gender.

If the baby clearly smiled, for example, everyone said the baby was smiling or happy.

SOURCES

Sex Differences: A Study of the Eye of the Beholder, Condry and Condry

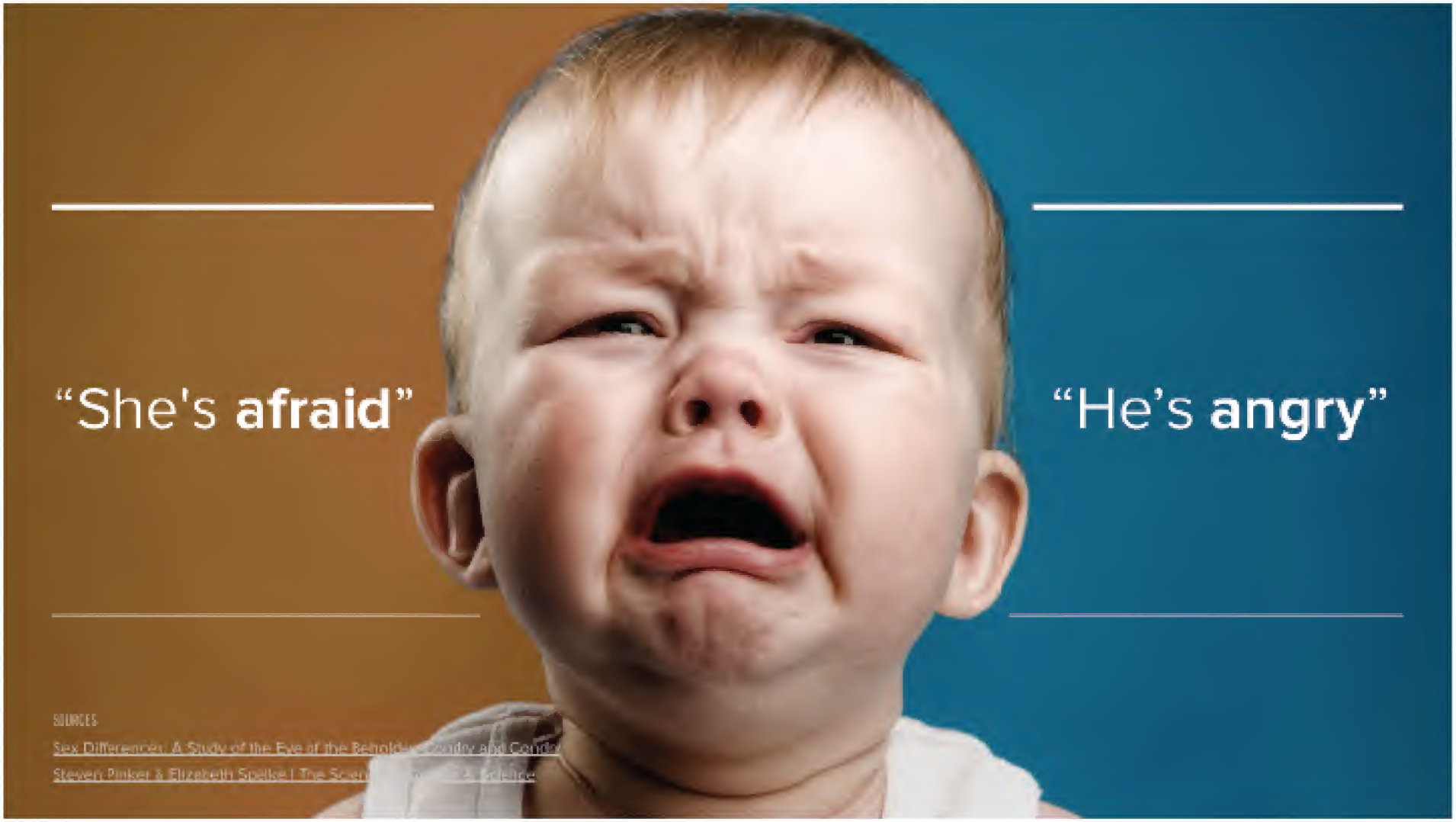
Steven Pinker & Elizabeth Spelke | The Science of Gender & Science

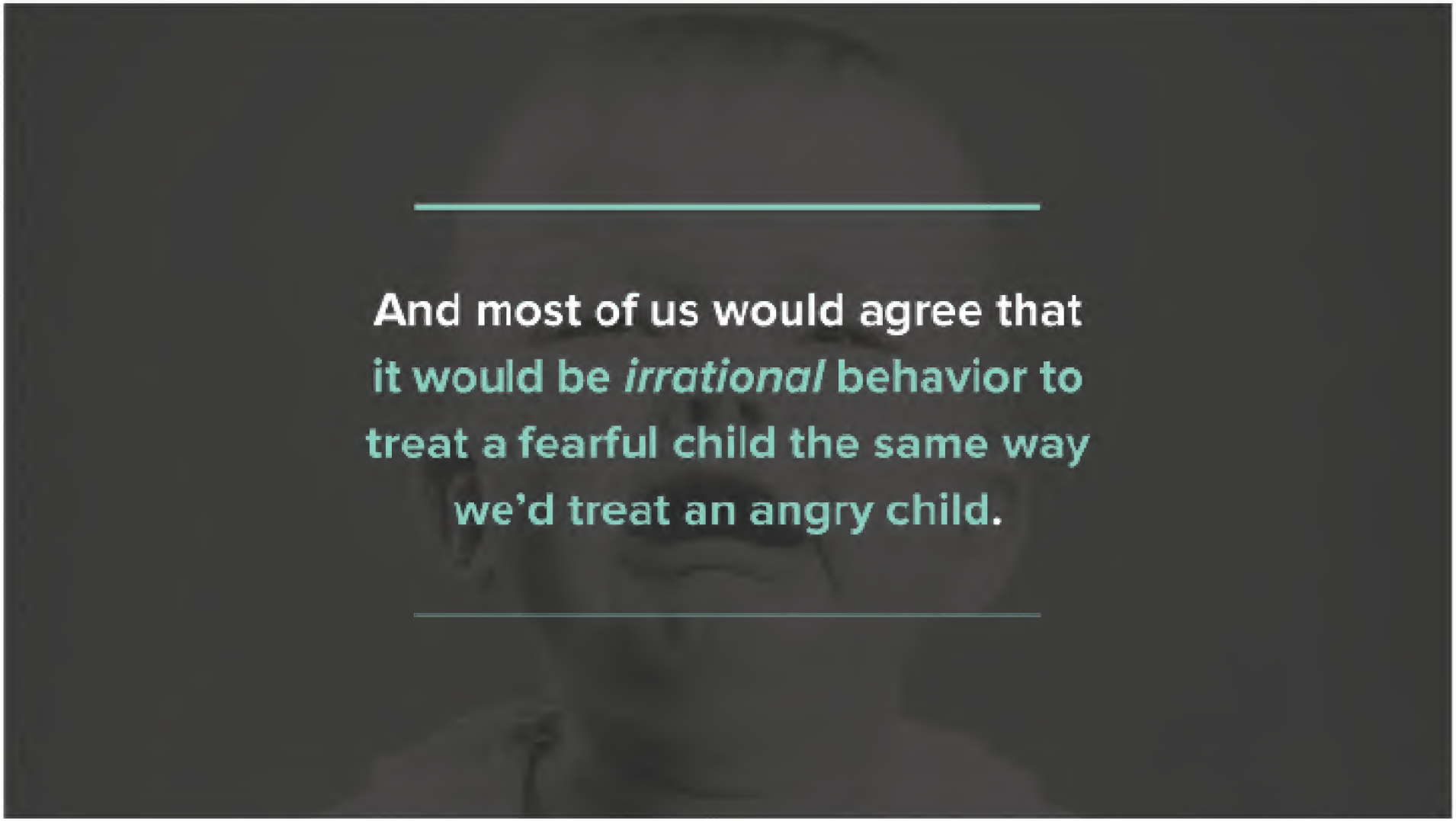
Then the babies played with a jack-in-the-box toy. When it suddenly popped up, the child was startled and jumped backward.

SOURCES

Sex Differences: A Study of the Eye of the Beholder, Connor and Connor

Steven Pinker & Elizabeth Spelke | The Science of Gender & Science


“She's **afraid**”
“He’s **angry**”
SOURCES
Sex Differences: A Study of the Eye of the Beholder
Steven Pinker & Elizabeth Spelke | The Scien


And most of us would agree that
it would be *irrational* behavior to
treat a fearful child the same way
we'd treat an angry child.

"If knowledge of a child's gender affects adults' perception of that child, then male and female children are going to elicit different reactions from the world, different patterns of encouragement. These perceptions matter, even in parents who are committed to treating sons and daughters alike."

— Elizabeth Spelke

SOURCES

Steven Pinker & Elizabeth Spelke | The Science of Gender & Science

Same child, same reaction,
different perception.

REPLICATION AND RELATED STUDIES

The Gender Stereotyping of Emotions, Plant, Hyde, Keltner, Devine

A comparison of observed and reported adult-infant interactions Effects of perceived sex, Culp, Cook, Housley

Adult perceptions of the infant as a function of gender labeling and observer gender, Delk, Madden, Livingston, Ryan

Surprising Smiles and Unsurprised Frowns: How Emotion and Status Influence Gender Categorization, Smith, Lefevre, Knol, Maes

Human perception drives virtually every facet of machine learning.

I propose we* make machine-learning intentionally human-centered and intervene for fairness.

* We = Humans. This isn't something any one company should be doing in isolation, but we're in a good position to start.

Tenets

Designing for fairness

01

Be Accountable

We can't take our hands off the steering wheel.

In rejecting the myth of neutral data, we are committing to be more conscious and conscientious.

01

Be Accountable

Present day
- ~~Robots~~
  - ~~Yada yada yada~~
    - ~~Future~~

Present day
- **Humans**
  - **Still humans**

02

Be Skeptical

Challenge assumptions at every turn.

We can't blindly rely on the systems that underpin conventional wisdom.

02

Be Skeptical

Journalism

Fake news is indistinguishable

Customer reviews

Male reviews skew average scores

Standardized tests

SAT scores don't predict grades

Medical science

Experiments over-recruit whites

Crime statistics

Racial profiling is real

03

Be Humane

Success metrics should bring out the best in human nature.

Standard engagement metrics confine people to whatever they've done *before*, rather than empowering what they're capable of doing *next*.

03

Be Humane

Time well spent

timewellspent.io

Learning and expression

Exploration and connection

User-defined goals

04

Be Humble

We don't *always* know better.

The tech industry is in love with "disruption", but frequently that means imposing a vision of the future *onto* users and expecting them to adapt.

04

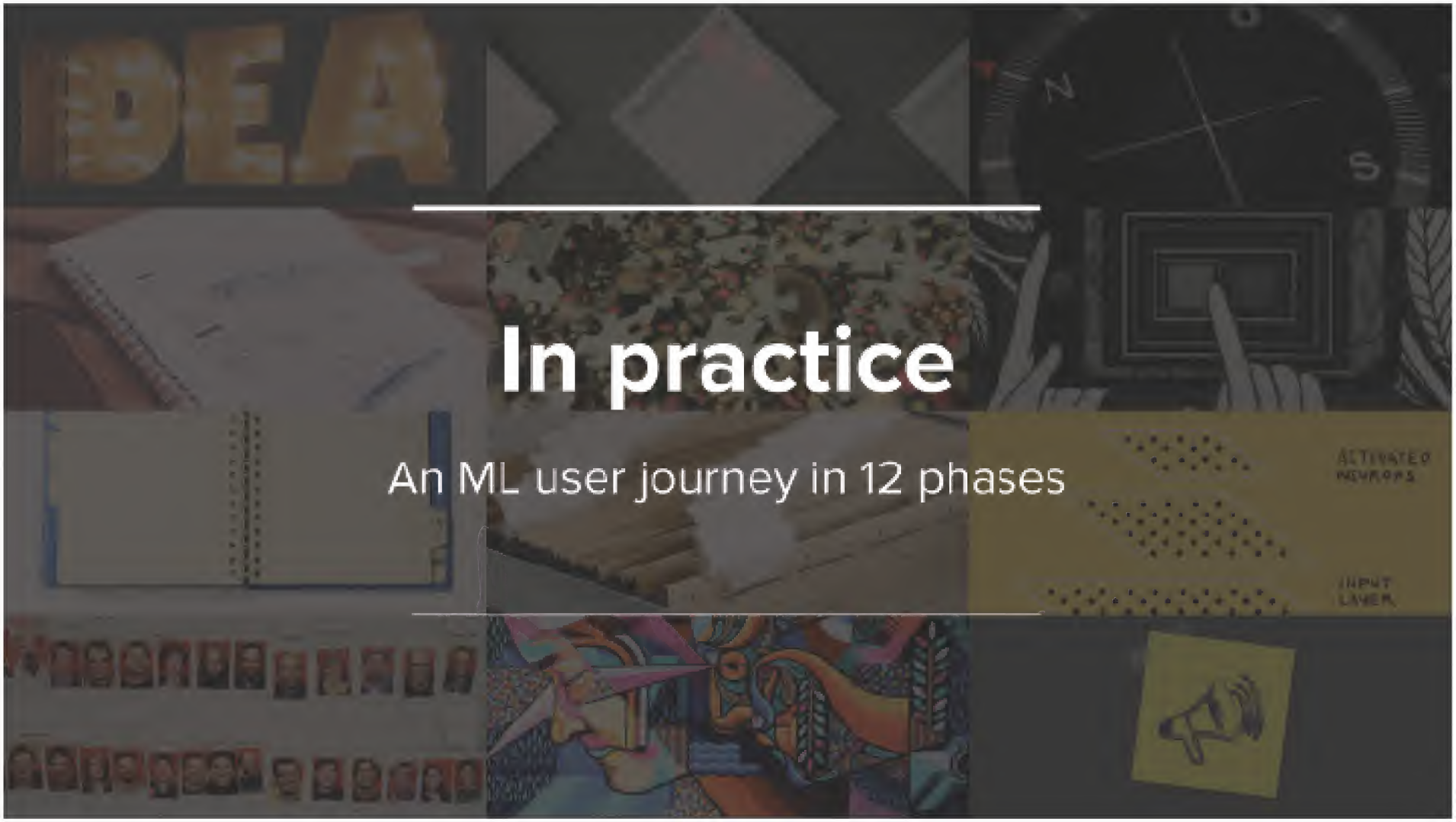
Be Humble

"There are two ways to get people used to automation. The soft and fuzzy way, ... just keep reassuring people until they're comfortable. And then there's the second way: let the humans ***take control when they need*** *to."*

—NPR

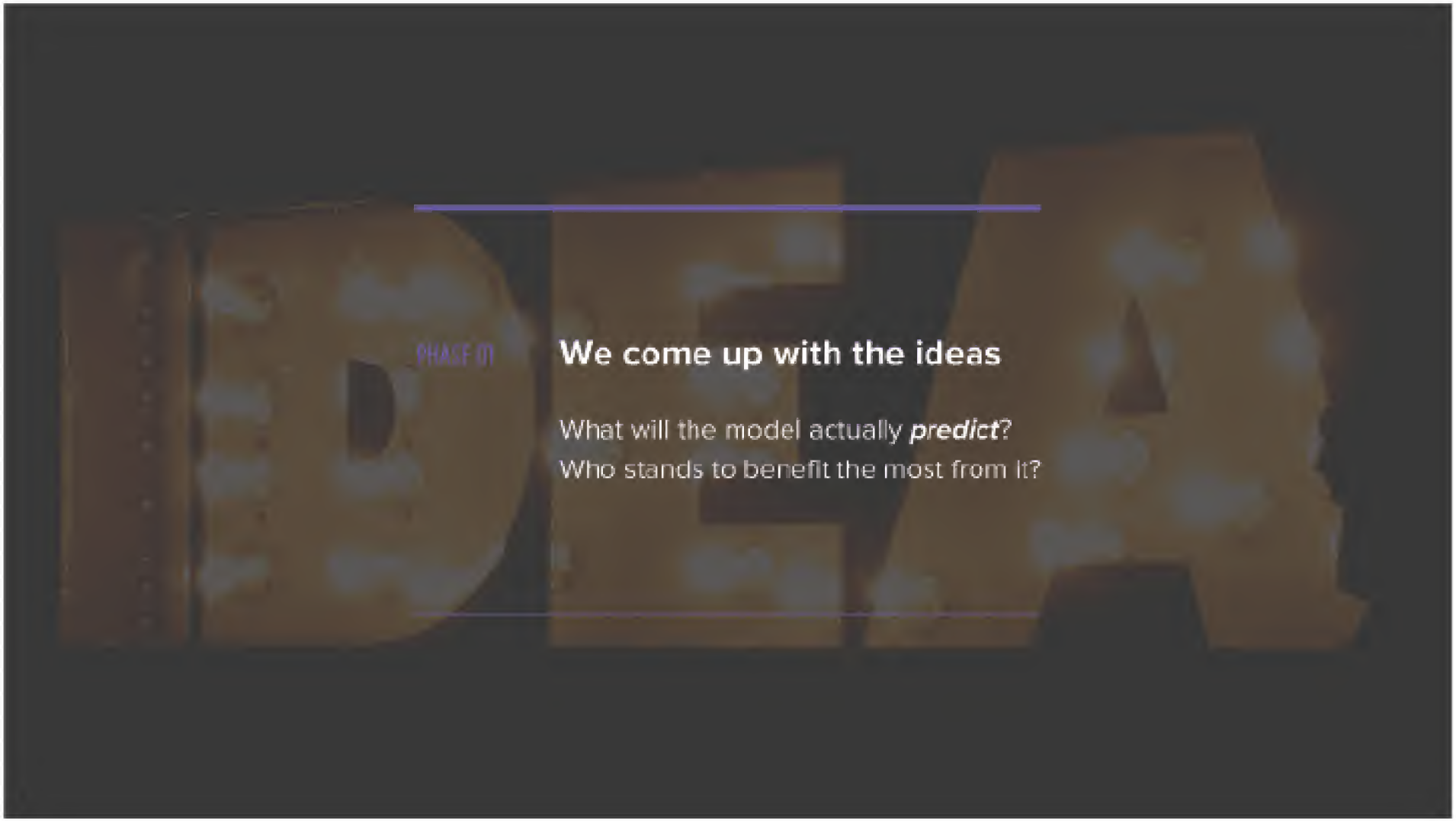
Augmenting → NOT → Automating

Supporting → NOT → Replacing



In practice

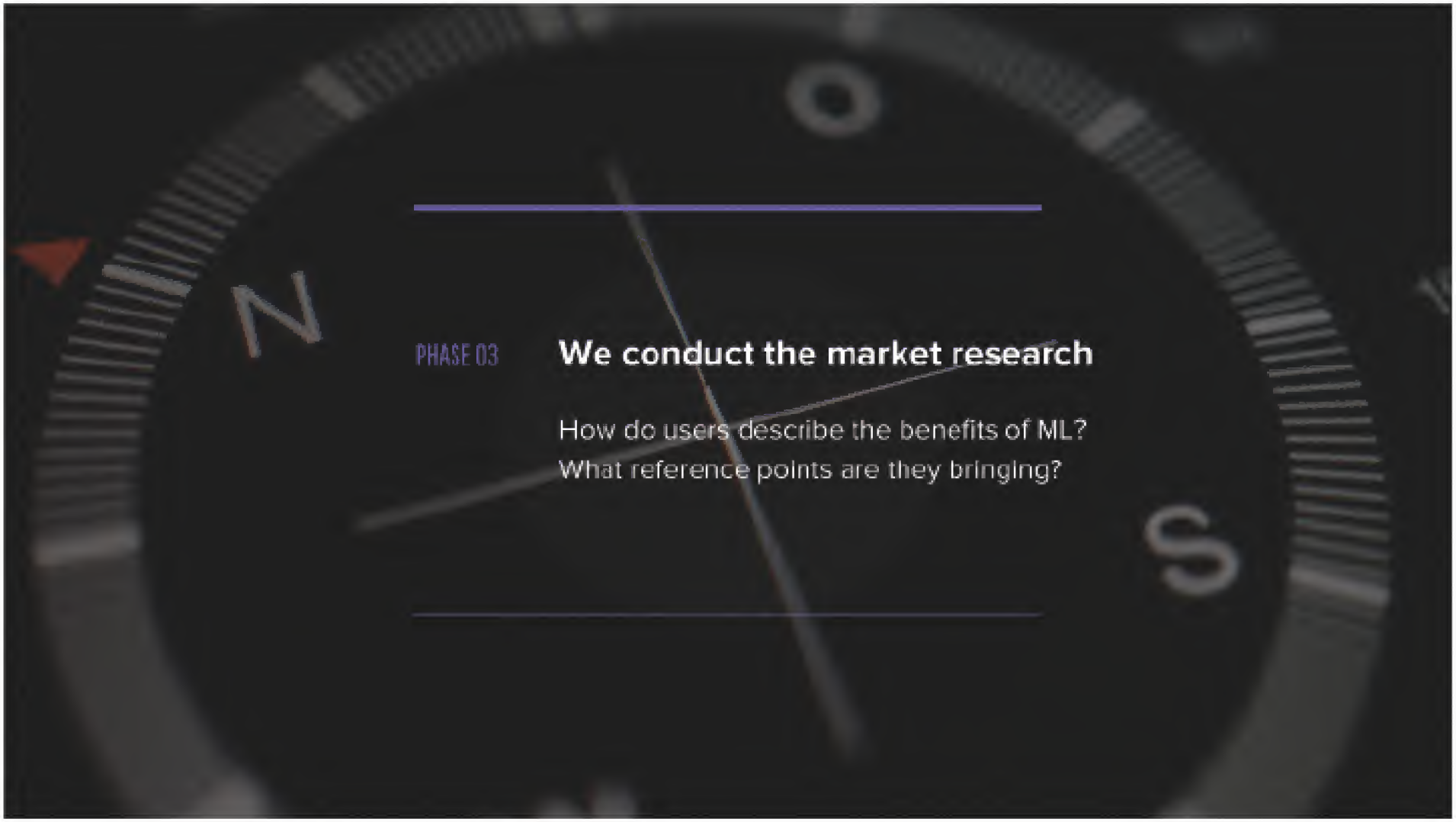
An ML user journey in 12 phases

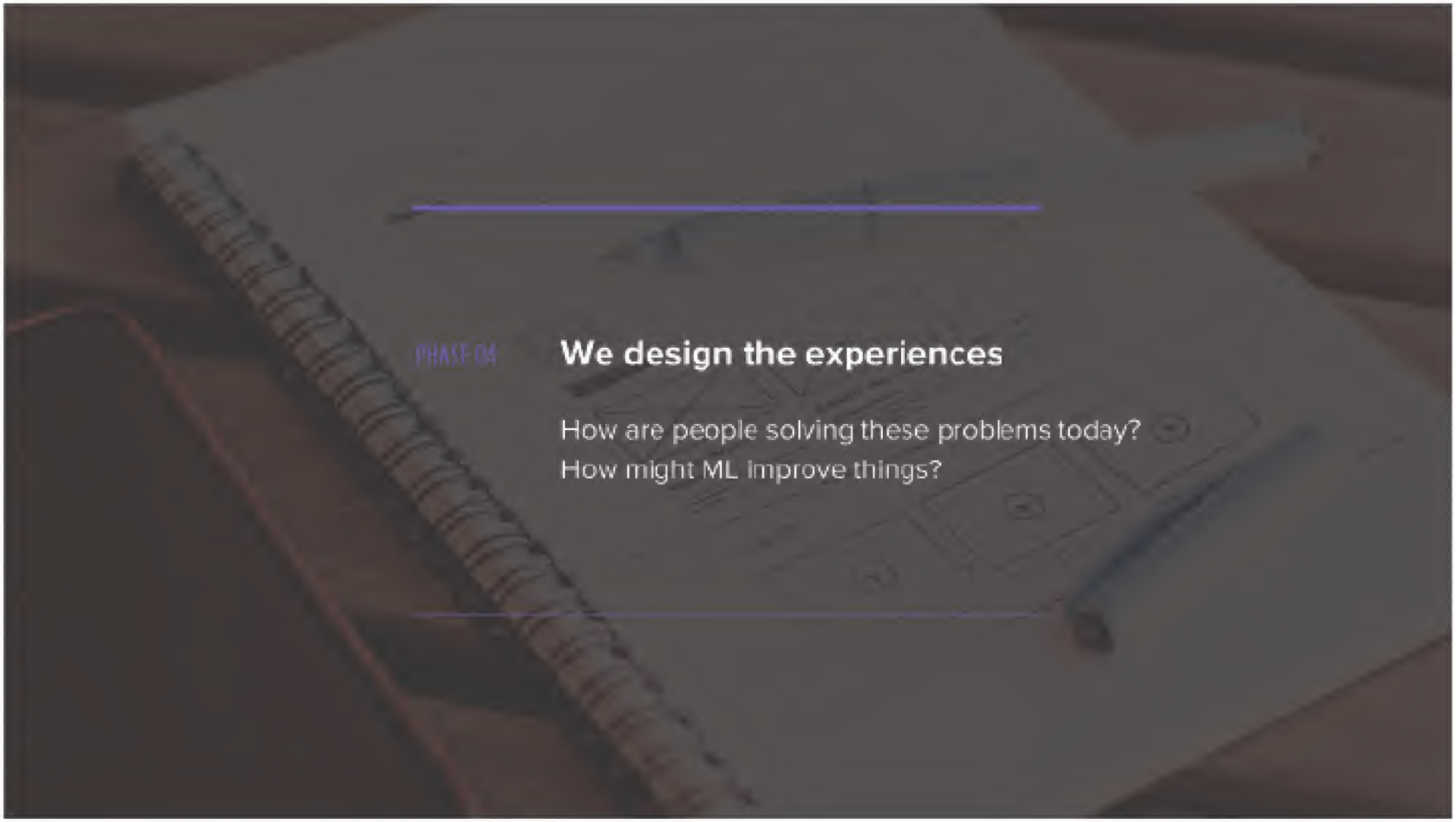

PHASE 01
We come up with the ideas
What will the model actually *predict*?
Who stands to benefit the most from it?

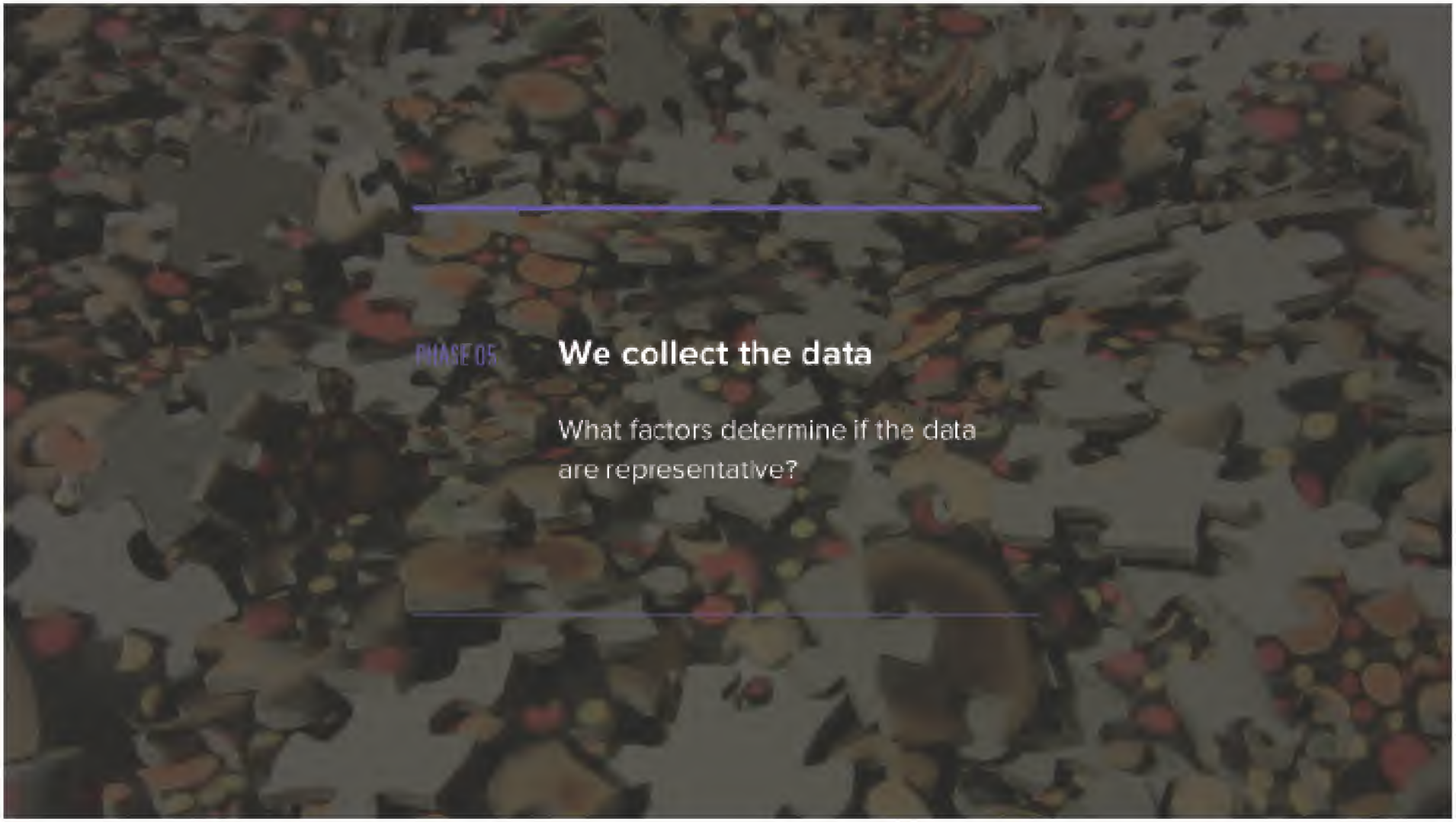
PHASE 02

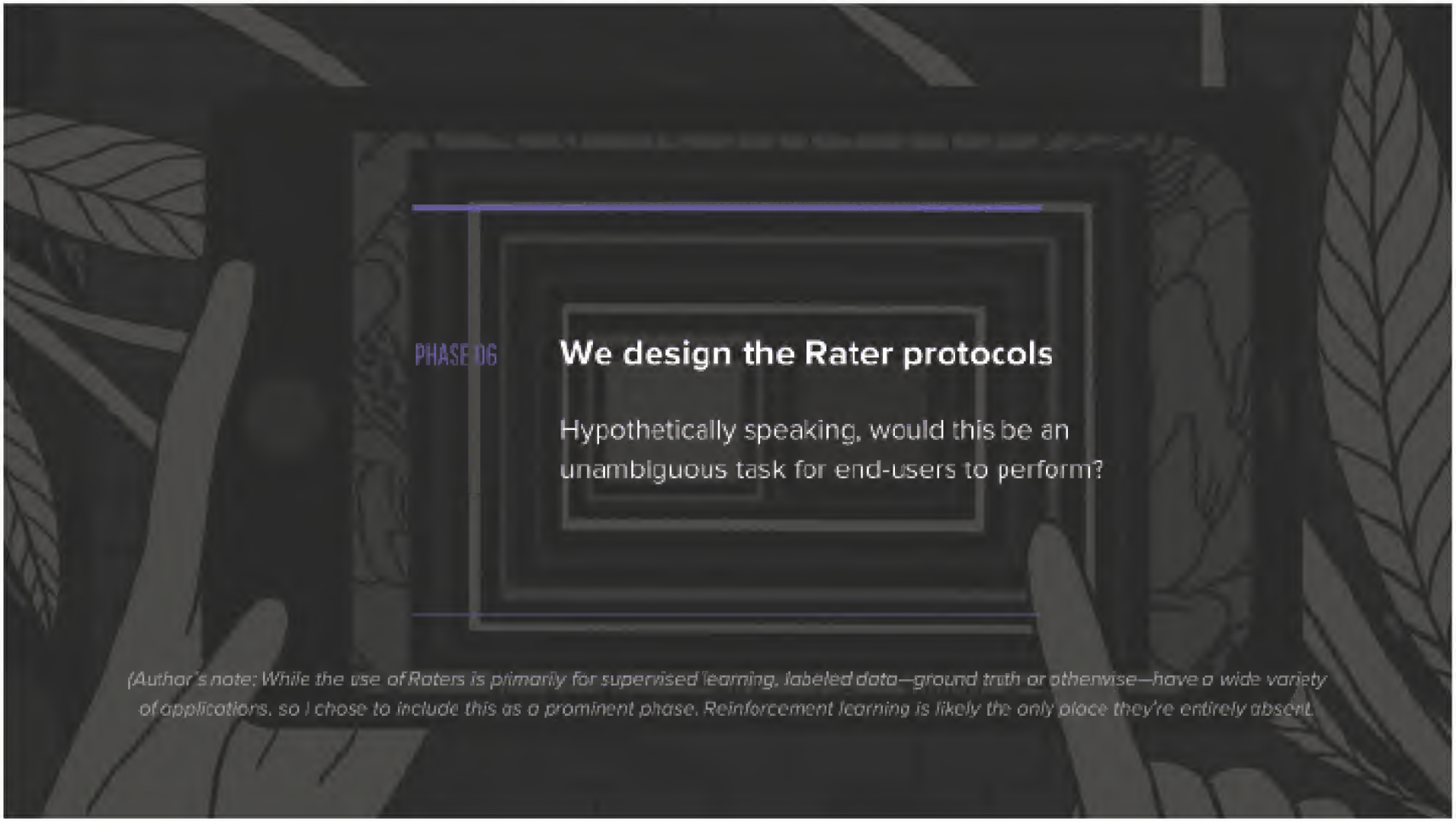
We define the ML strategy

Why would heuristics be less effective than a machine-learned model?

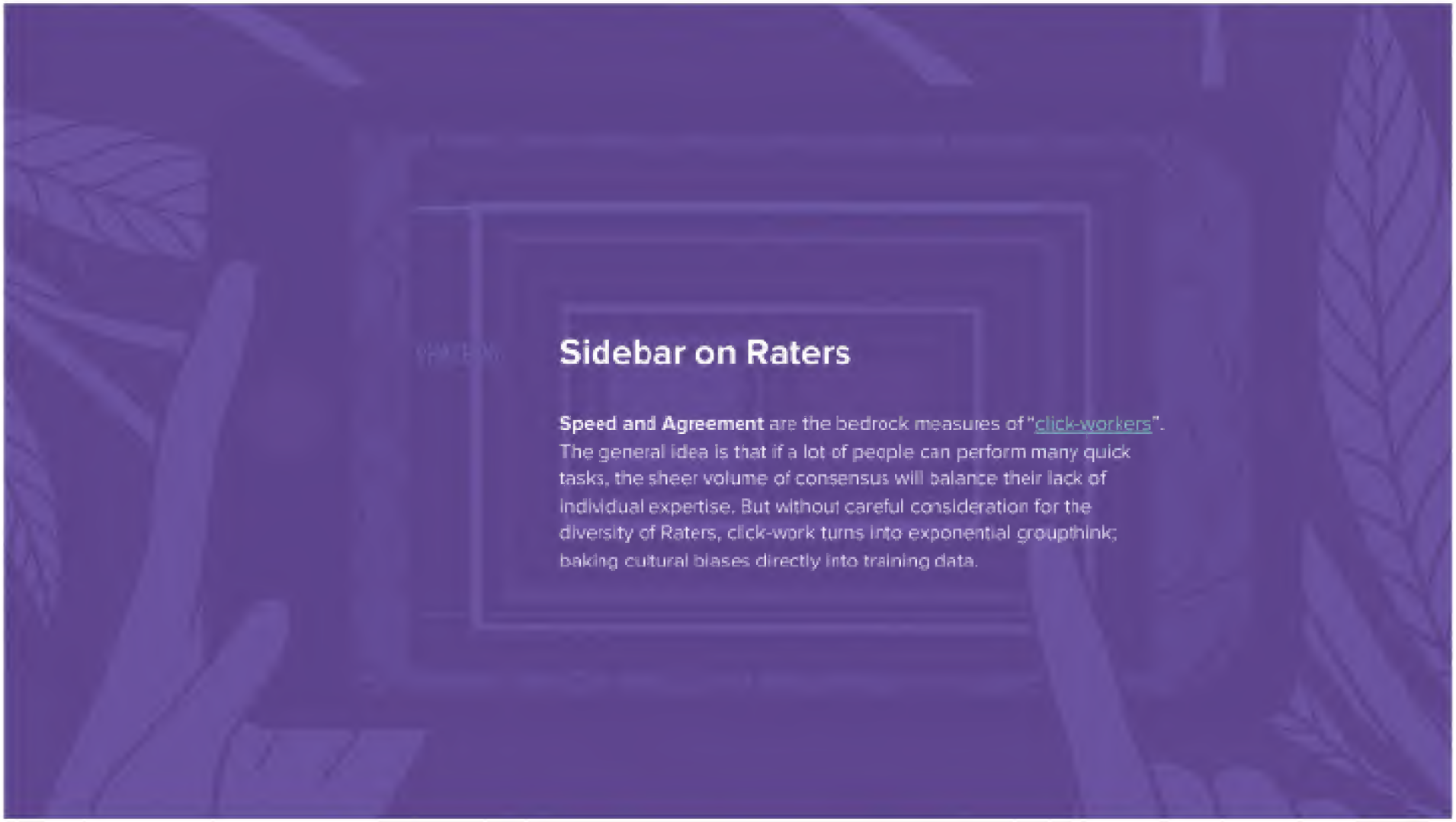

PHASE 03
We conduct the market research
How do users describe the benefits of ML?
What reference points are they bringing?


PHASE 04
We design the experiences
How are people solving these problems today?
How might ML improve things?


PHASE 05
We collect the data
What factors determine if the data are representative?

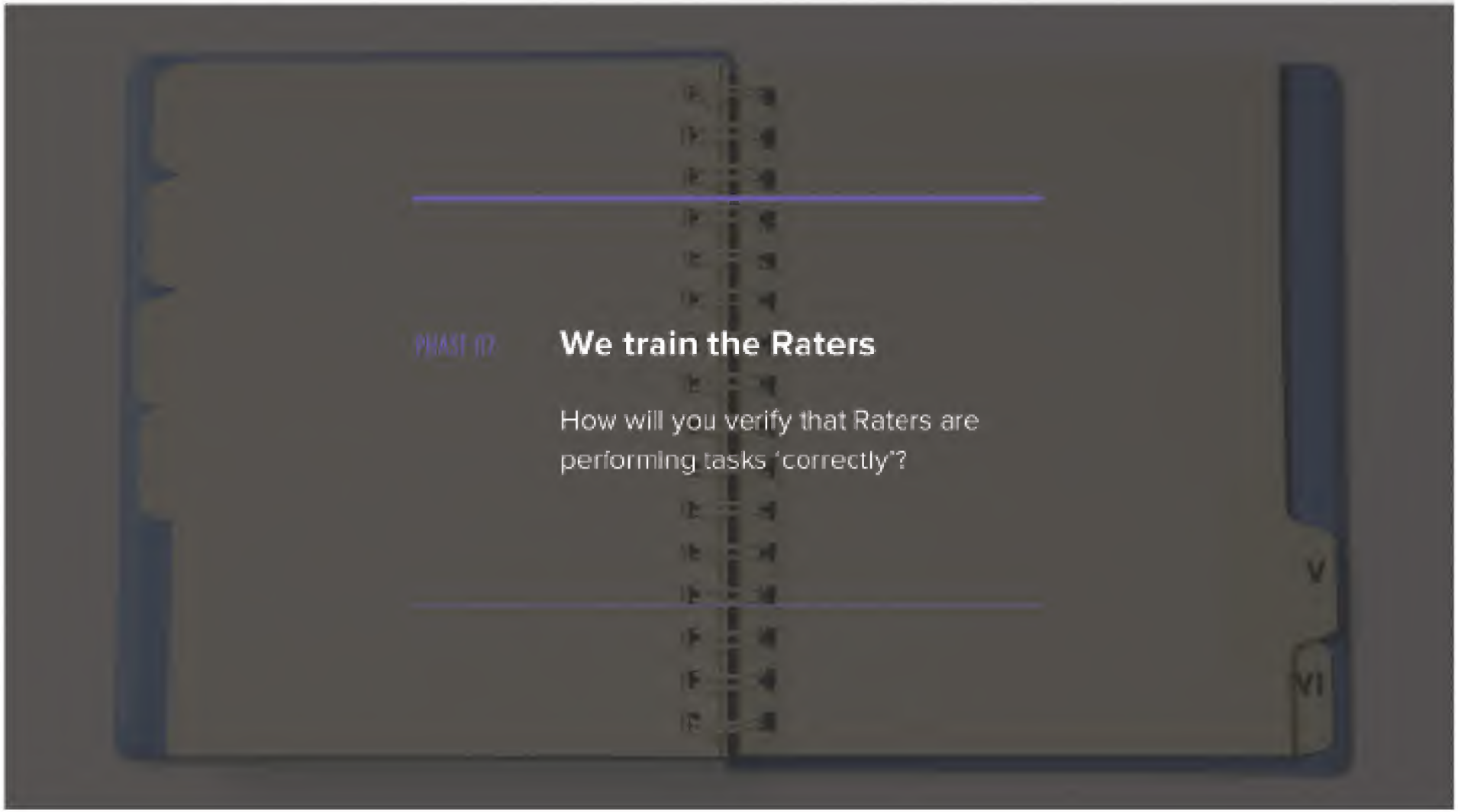


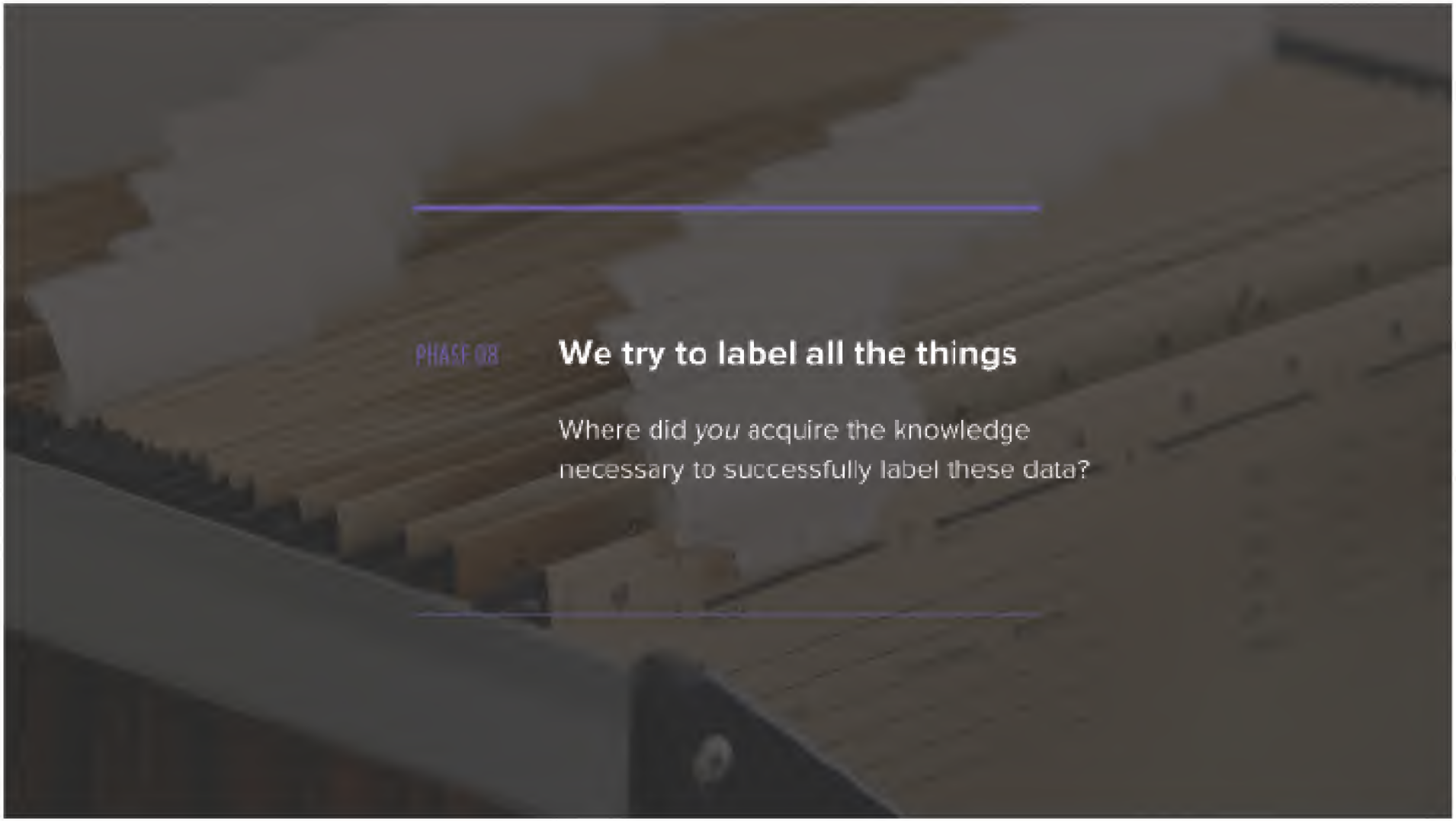
(Author's note: While the use of Raters is primarily for supervised learning, labeled data—ground truth or otherwise—have a wide variety of applications, so I chose to include this as a prominent phase. Reinforcement learning is likely the only place they're entirely absent.

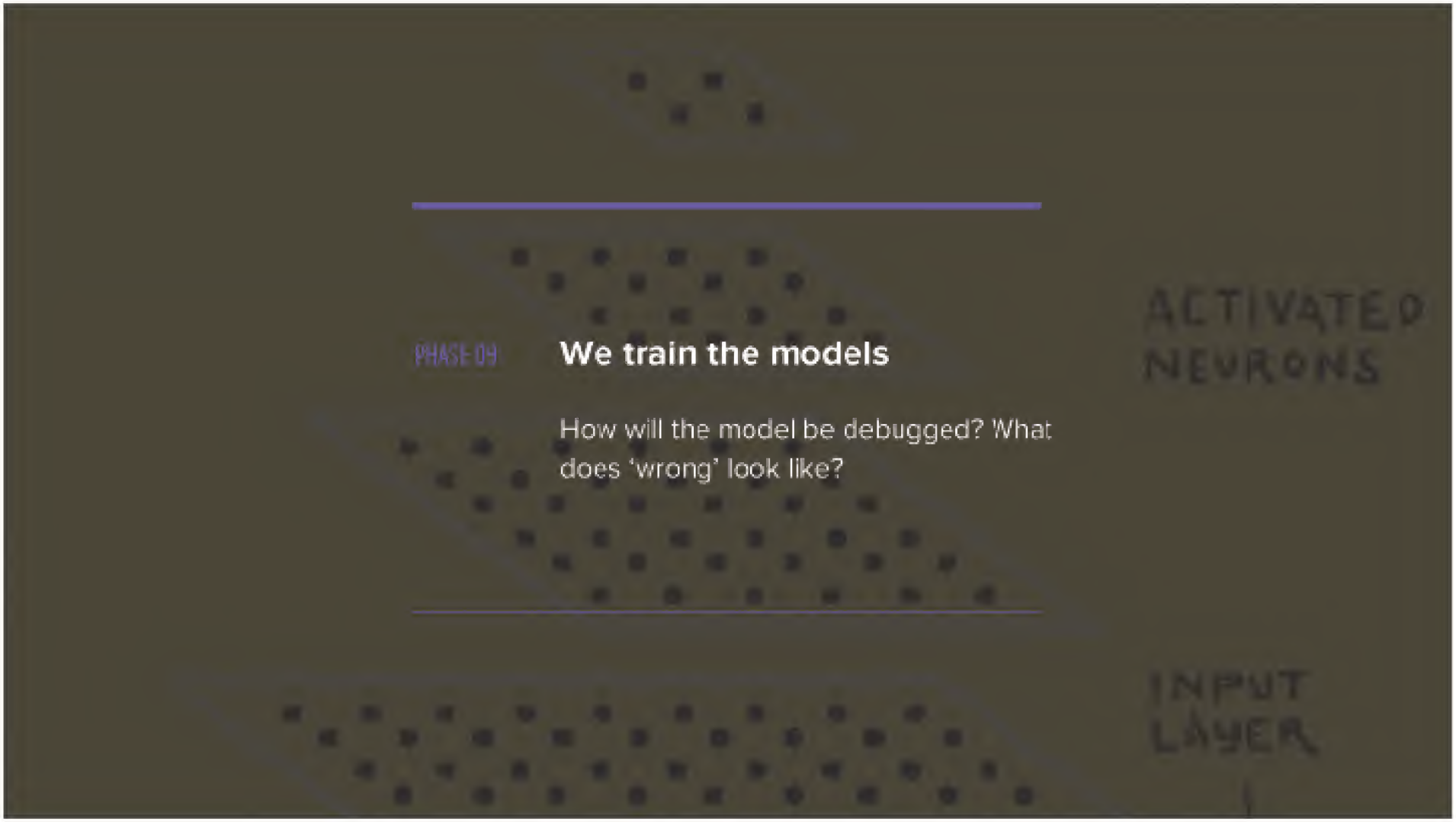


Sidebar on Raters

Speed and Agreement are the bedrock measures of "click-workers". The general idea is that if a lot of people can perform many quick tasks, the sheer volume of consensus will balance their lack of individual expertise. But without careful consideration for the diversity of Raters, click-work turns into exponential groupthink; baking cultural biases directly into training data.


PHASE 02
We train the Raters
How will you verify that Raters are performing tasks ‘correctly’?
V
VI

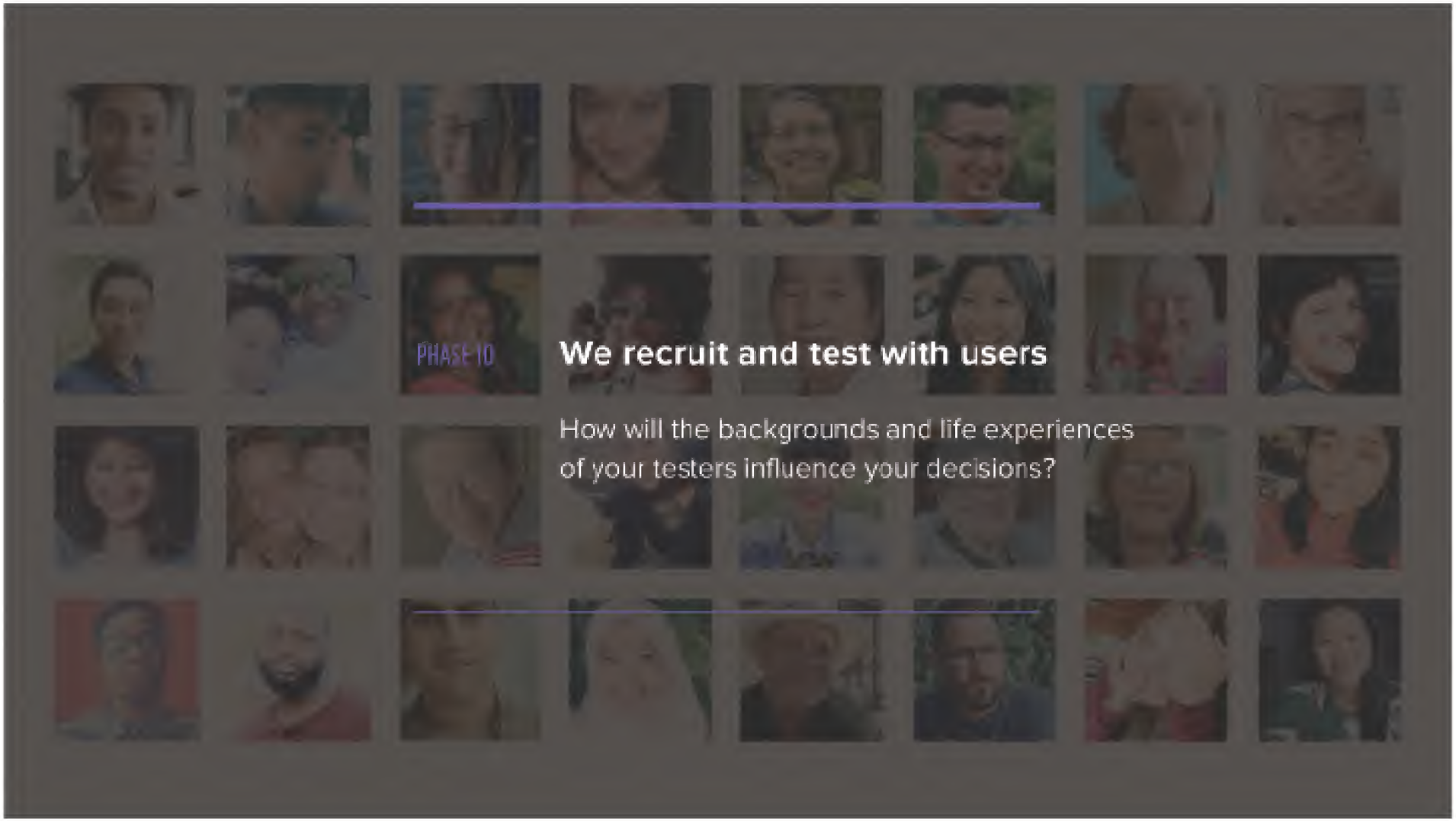

PHASE 08
We try to label all the things
Where did *you* acquire the knowledge necessary to successfully label these data?

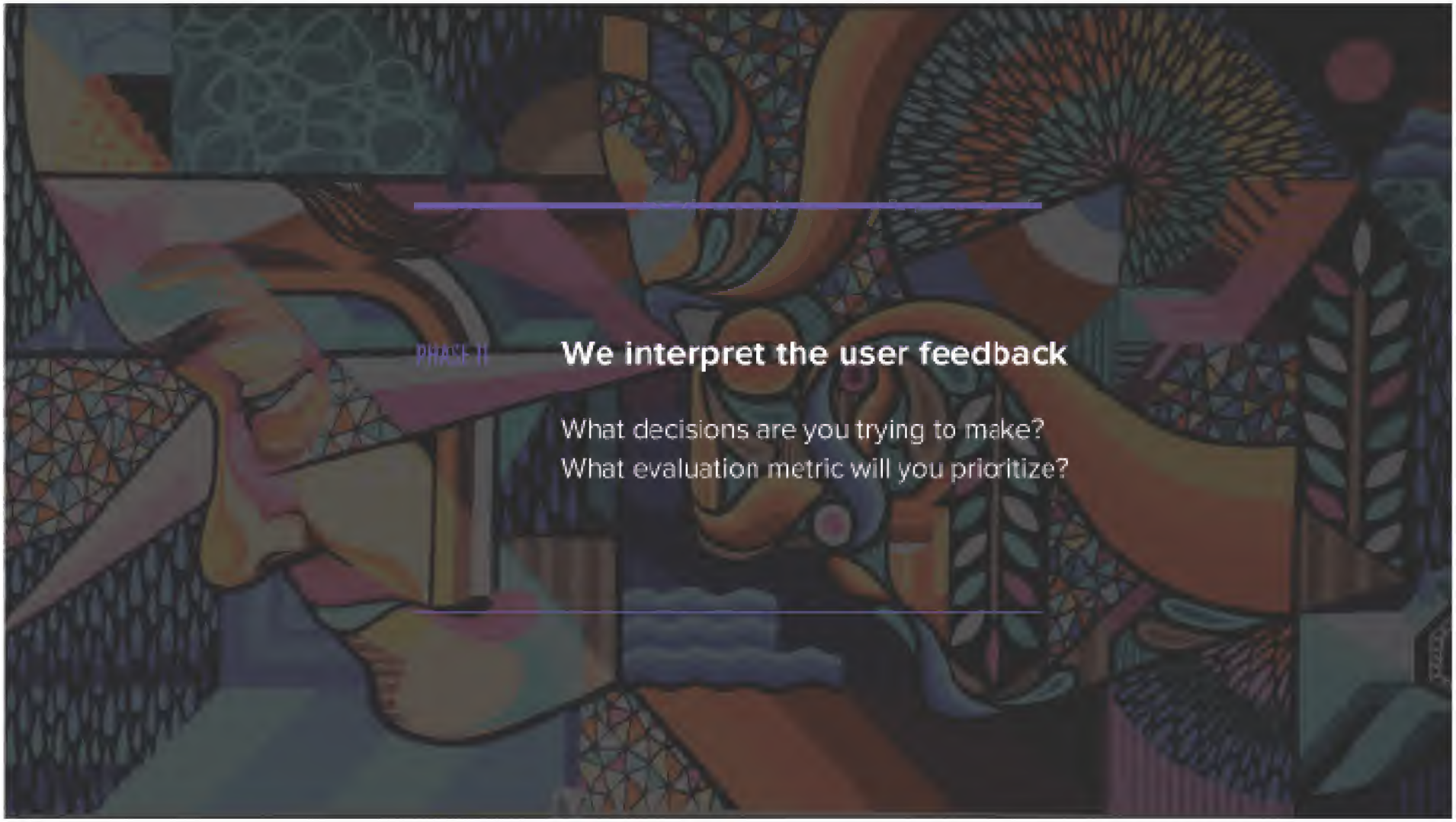


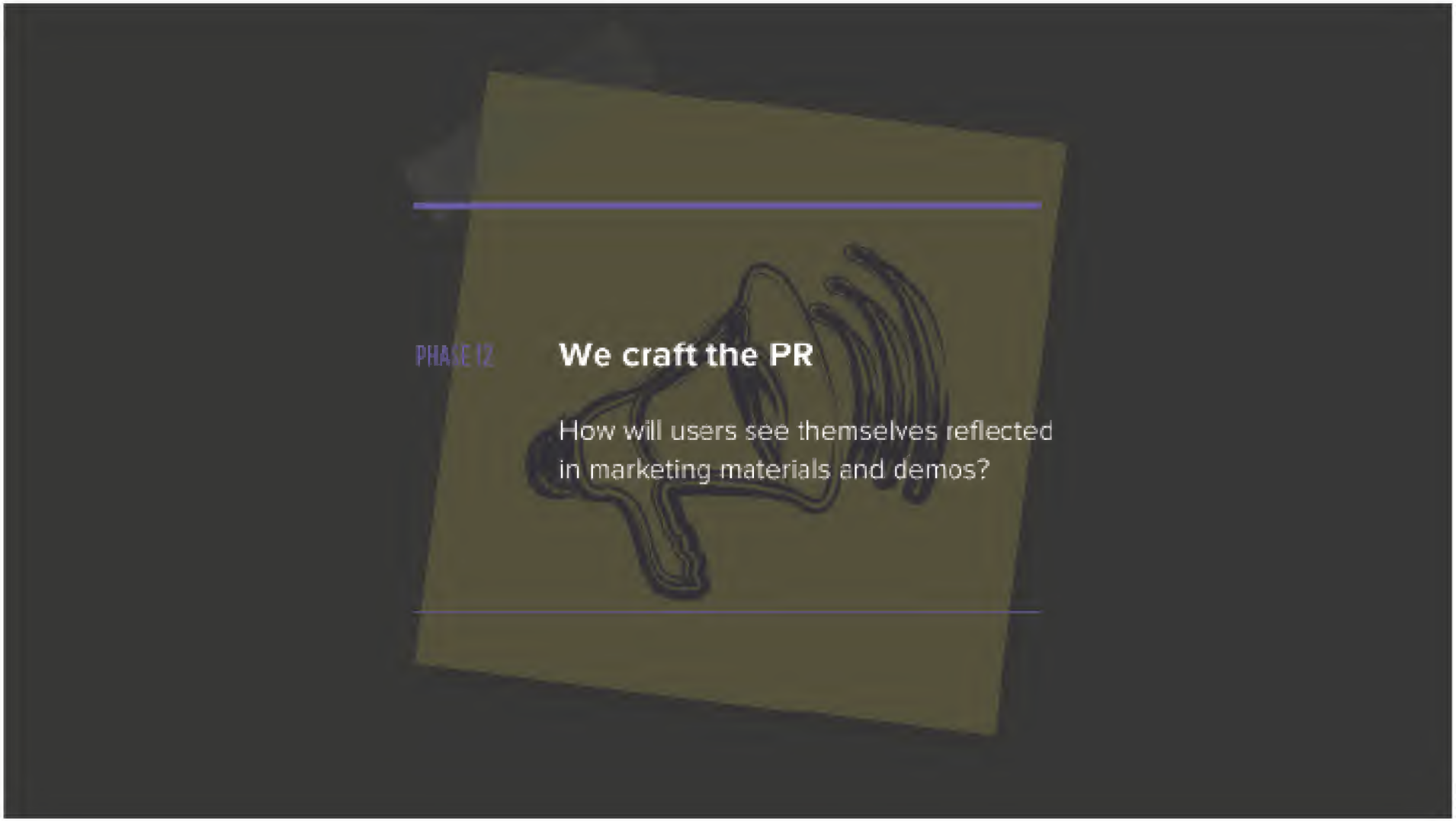
PHASE 04

We train the models

How will the model be debugged? What does 'wrong' look like?


PHASE 10
We recruit and test with users
How will the backgrounds and life experiences
of your testers influence your decisions?


PHASE II
We interpret the user feedback
What decisions are you trying to make?
What evaluation metric will you prioritize?


PHASE 12
We craft the PR
How will users see themselves reflected in marketing materials and demos?

Finally, I'd like to offer **3 human-centered diagnostics** when designing with ML.

If you're finding it tricky to answer these, it might be a signal to slow down and take a closer look.

01

If a human were to perform this task, what would ‘appropriate’ social behavior look like?

What interpersonal cues might be relevant that are missing from your input or interface? E.g. body language, tone of voice.

Take **autocomplete** for example: In what context would it be acceptable to finish another person's sentence before they've stopped talking?

Bear in mind that no one culture is capable of representing universal norms, especially for social interactions, so we need to be mindful of our intuitions. The psychological effects of algorithmic discrimination likely mirror those of social discrimination.

02

If a human were to perform this task, might we call it an expression of their personality?

Tasks that have unambiguous utility, are perceived as repetitive or boring by users, and/or benefit from super fast response times are ideal candidates for ML augmentation.

Grey area: Suggesting a reply in an email or SMS likely has a priming effect: impacting the user's response even if they don't use it. And the fact that suggestions are even offered may lead the recipient to question the authenticity of the sentiment.

Augmentation: The goals of a **self-driving car** are unambiguous, and the benefits of a computer's superhuman reaction time offer objective utility. No one (hopefully) would say a driver got into an accident because they were expressing themselves.

03

Who are you?

... OK, now what do your data teach you about **everyone *else*?**

Our traits don't necessarily define us, but it's foolish to pretend we don't see them. By taking the potentially uncomfortable step of inventorying these traits—physical, social, cognitive, and otherwise—we're **getting proximate to those who are reminded of their differentness every time a 'default' is invoked in day-to-day life.**

I am...

White	Male
A parent	Affluent
Visually impaired	Physically capable
Agnostic	Culturally Jewish
Insured	A homeowner
Urban-dwelling	Not college educated
In my 30's	Married
Cisgender	Heterosexual
99% percentile height	In good mental health
A speaker of "standard" U.S. english	

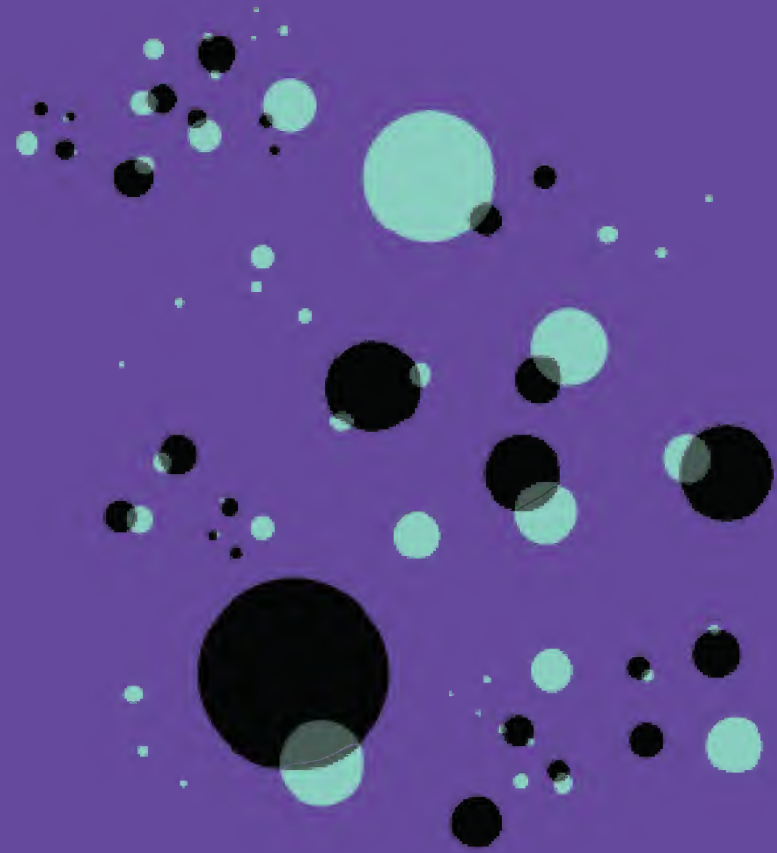
03

Because hopefully that can help your world look a bit **less like this...**

03

Because hopefully that can help your world look a bit **less like this**...

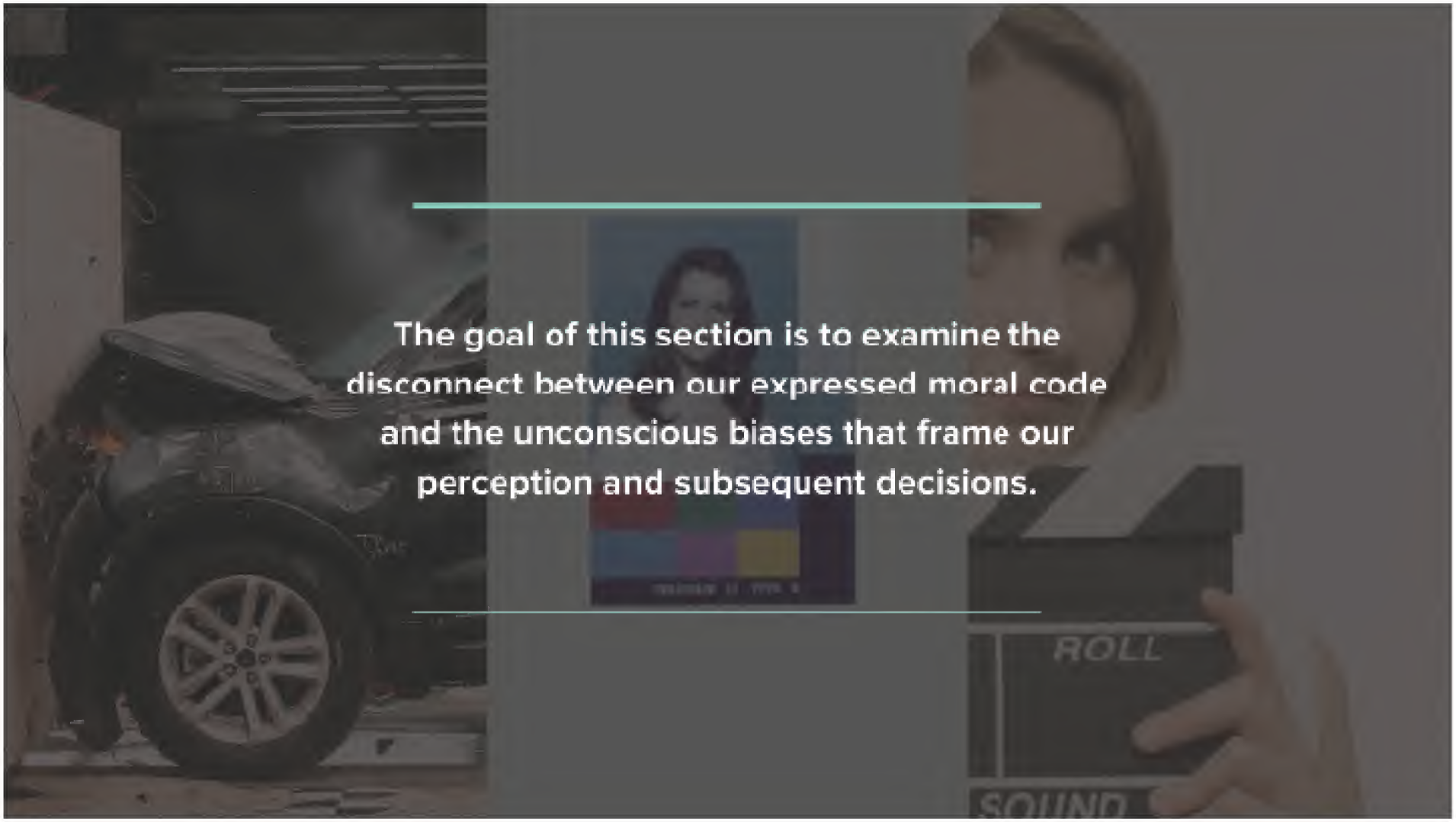
...and a lot **more like this**



Thank you

Dive deeper at go/ml-fairness

Appendix



The goal of this section is to examine the disconnect between our expressed moral code and the unconscious biases that frame our perception and subsequent decisions.

01 Gender roles in film

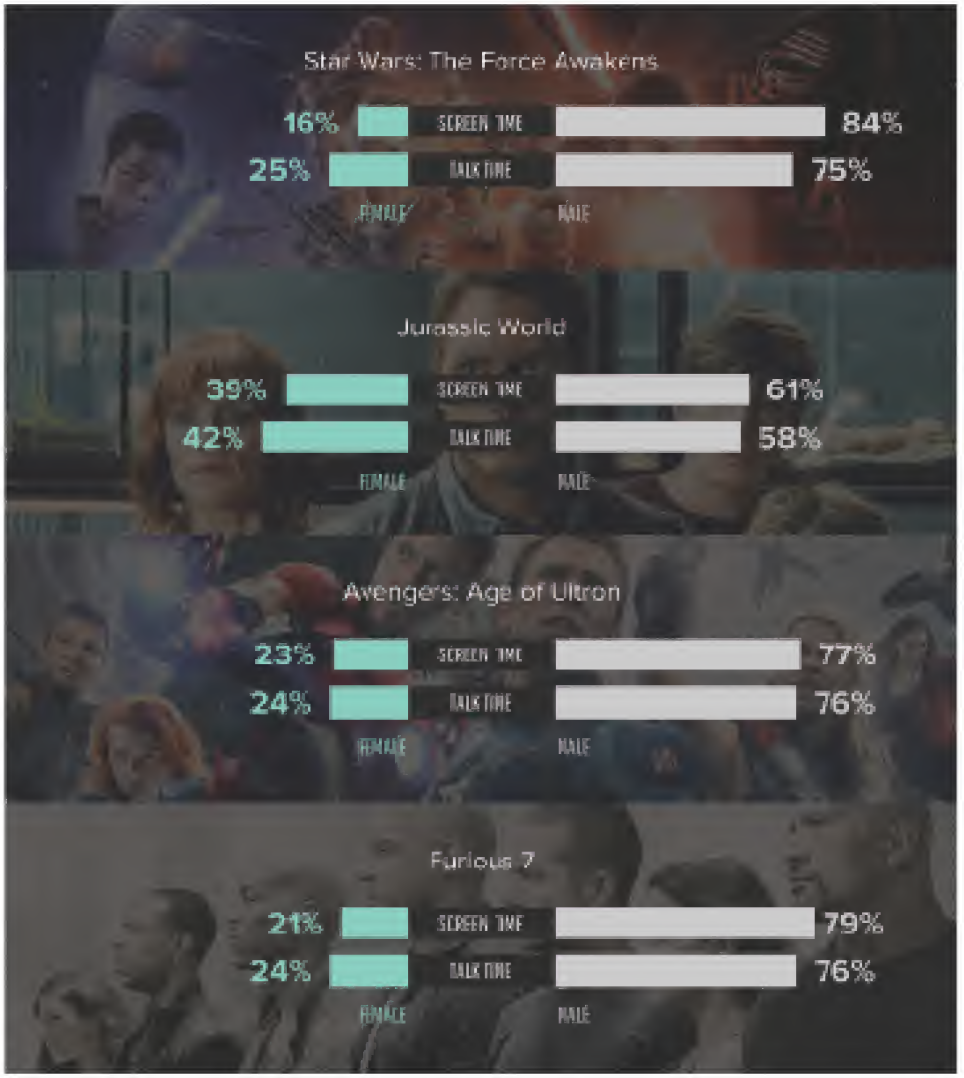
In 2015, women shared top billing with men in the top four grossing live-action films in the U.S.

How much were female characters seen and heard compared to men in those films?

GENDER ROLES IN FILM

In 2015, women shared top billing with men in the top four grossing live-action films in the U.S.

How much were **female characters seen and heard compared to men** in those films?



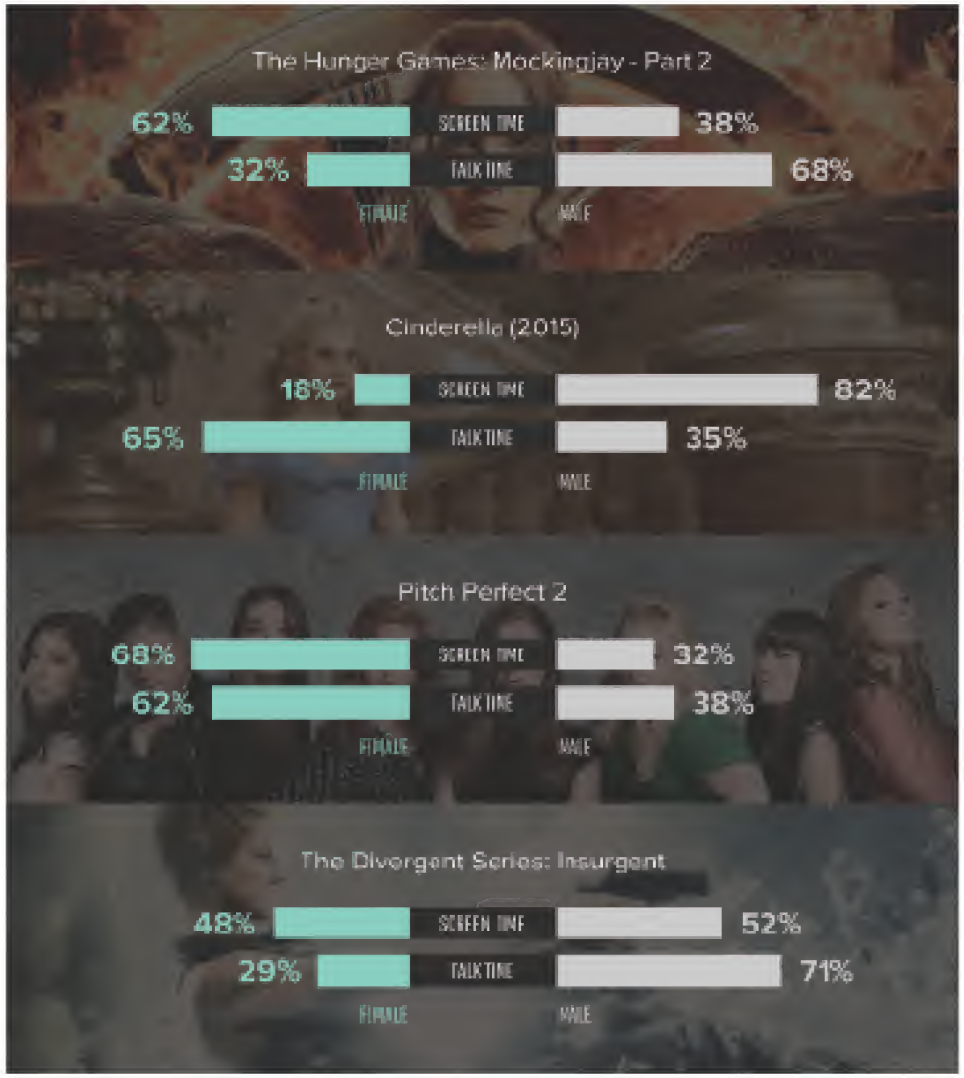
SOURCE
Geena Davis Institute on Gender in Media

The numbers improved slightly for the top grossing films featuring female leads.

But still contain some surprisingly disproportionate numbers.

The numbers improved slightly for the top grossing films featuring **female leads**.

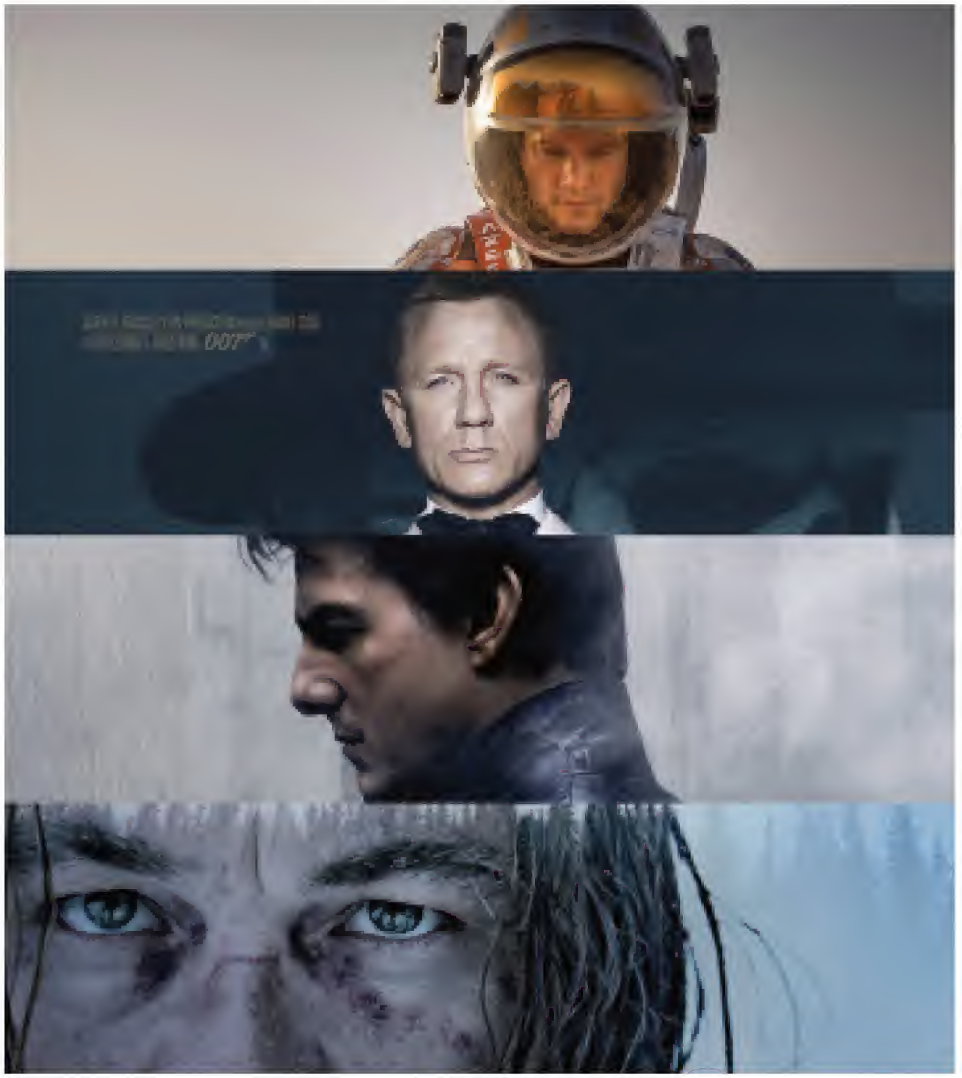
But still contain some surprisingly disproportionate numbers.



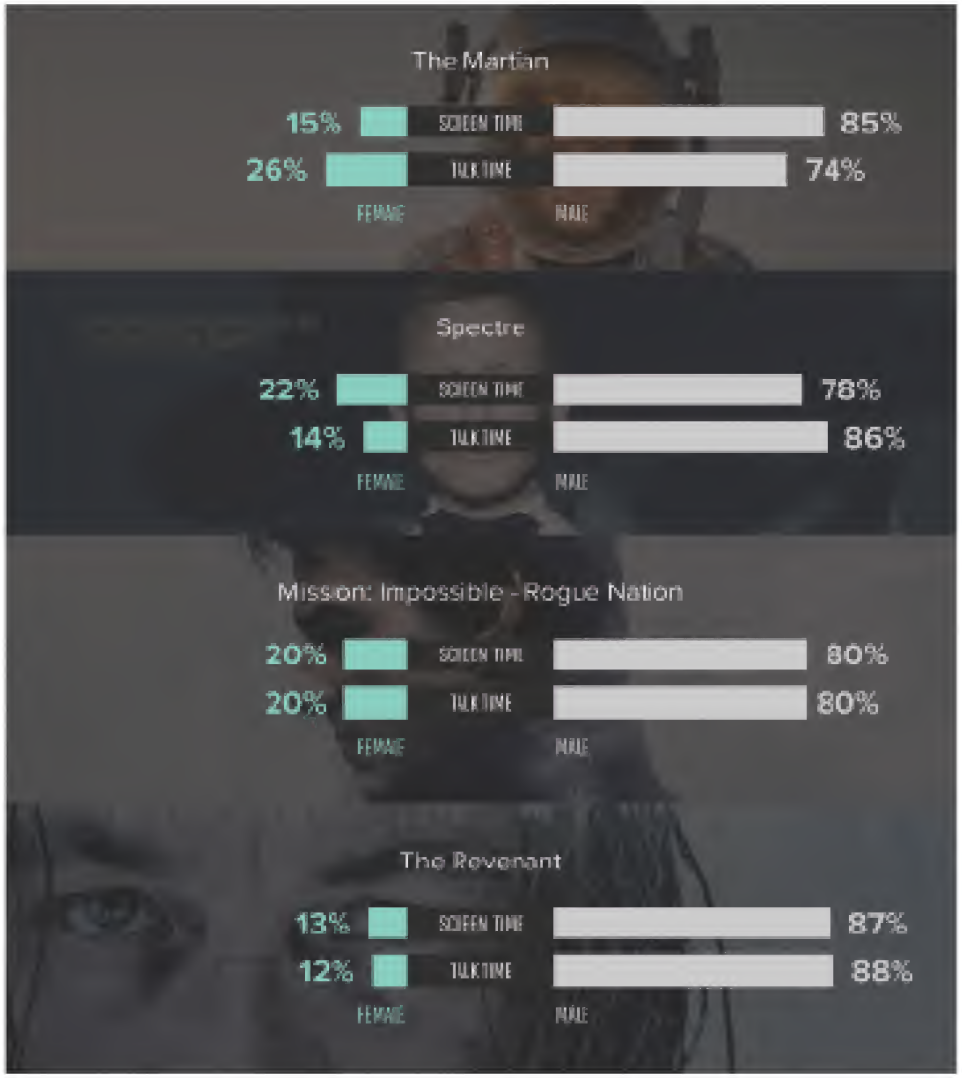
SOURCE

Geena Davis Institute on Gender in Media

But the gap increased substantially for the top grossing films featuring male leads.

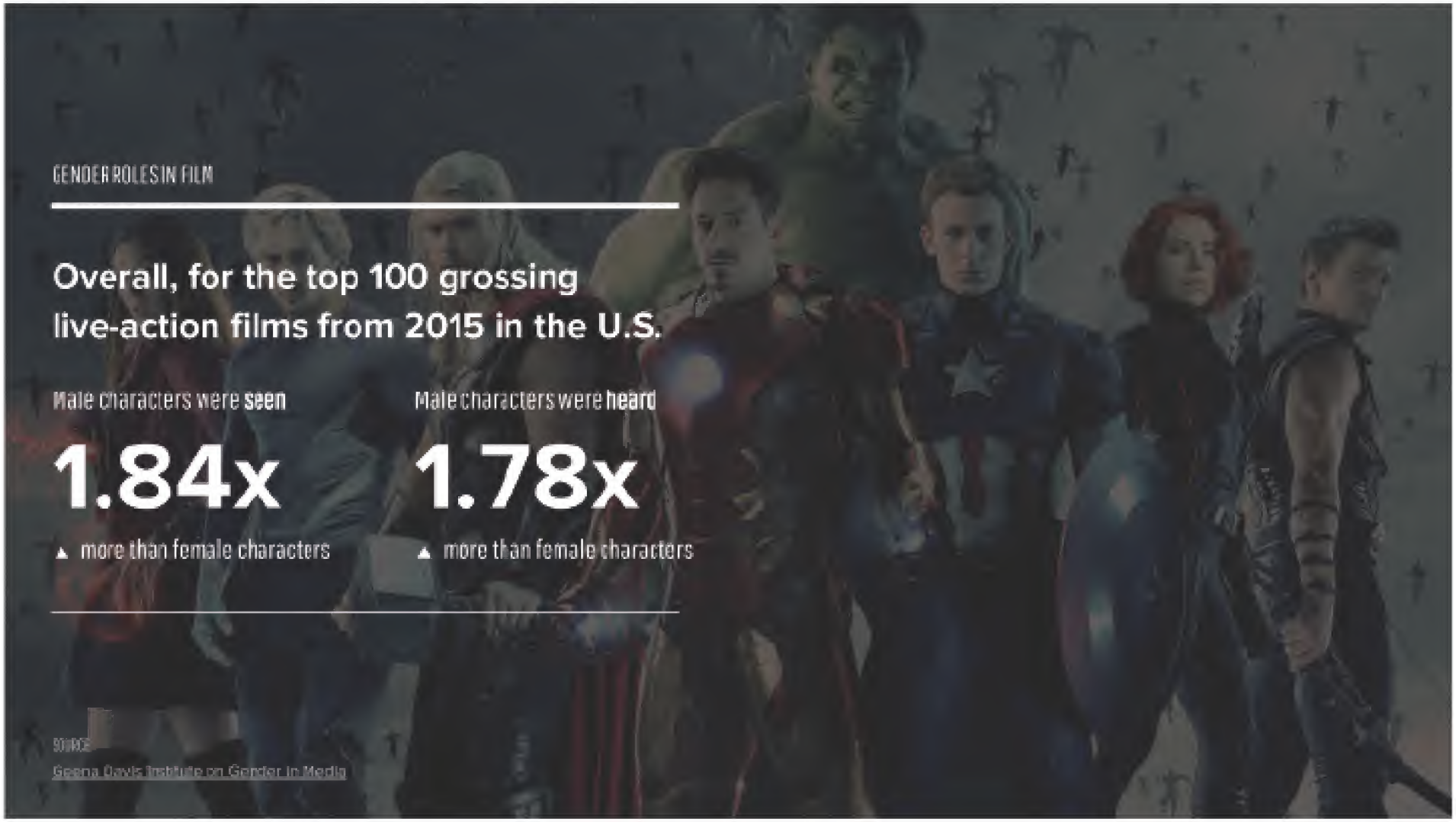


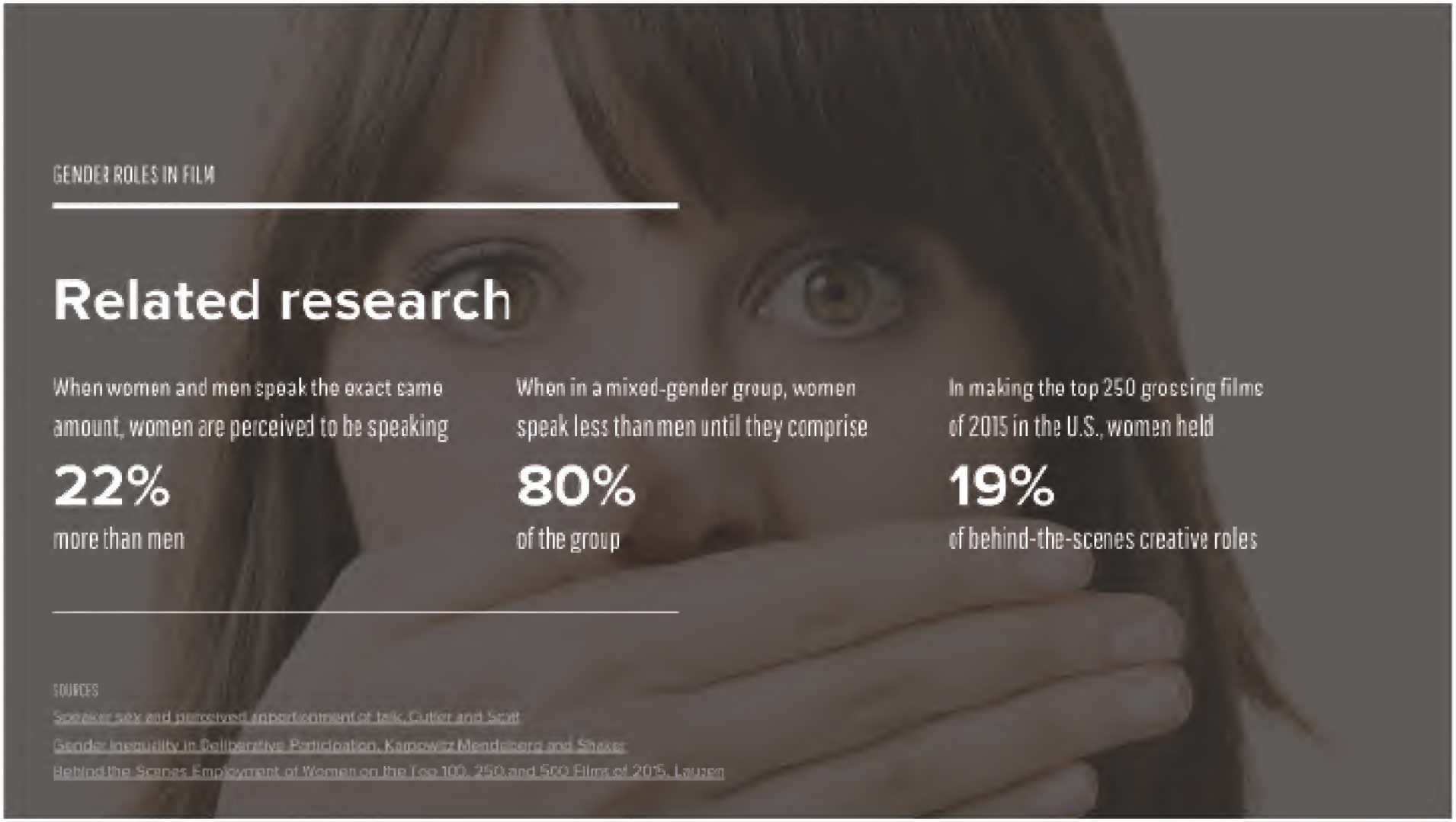
But the gap increased substantially for the top grossing films featuring **male leads**.



SOURCE

Geena Davis Institute on Gender in Media


GENDER ROLES IN FILM
Overall, for the top 100 grossing live-action films from 2015 in the U.S.
Male characters were seen
1.84x
▲ more than female characters
Male characters were heard
1.78x
▲ more than female characters
SOURCE
Geena Davis Institute on Gender in Media



GENDER ROLES IN FILM

Related research

When women and men speak the exact same amount, women are perceived to be speaking

22%

more than men

When in a mixed-gender group, women speak less than men until they comprise

80%

of the group

In making the top 250 grossing films of 2015 in the U.S., women held

19%

of behind-the-scenes creative roles

SOURCES

Speaker sex and perceived apportionment of talk, Cutler and Scott

Gender Inequality in Deliberative Participation, Karpowitz Mendelberg and Shaker

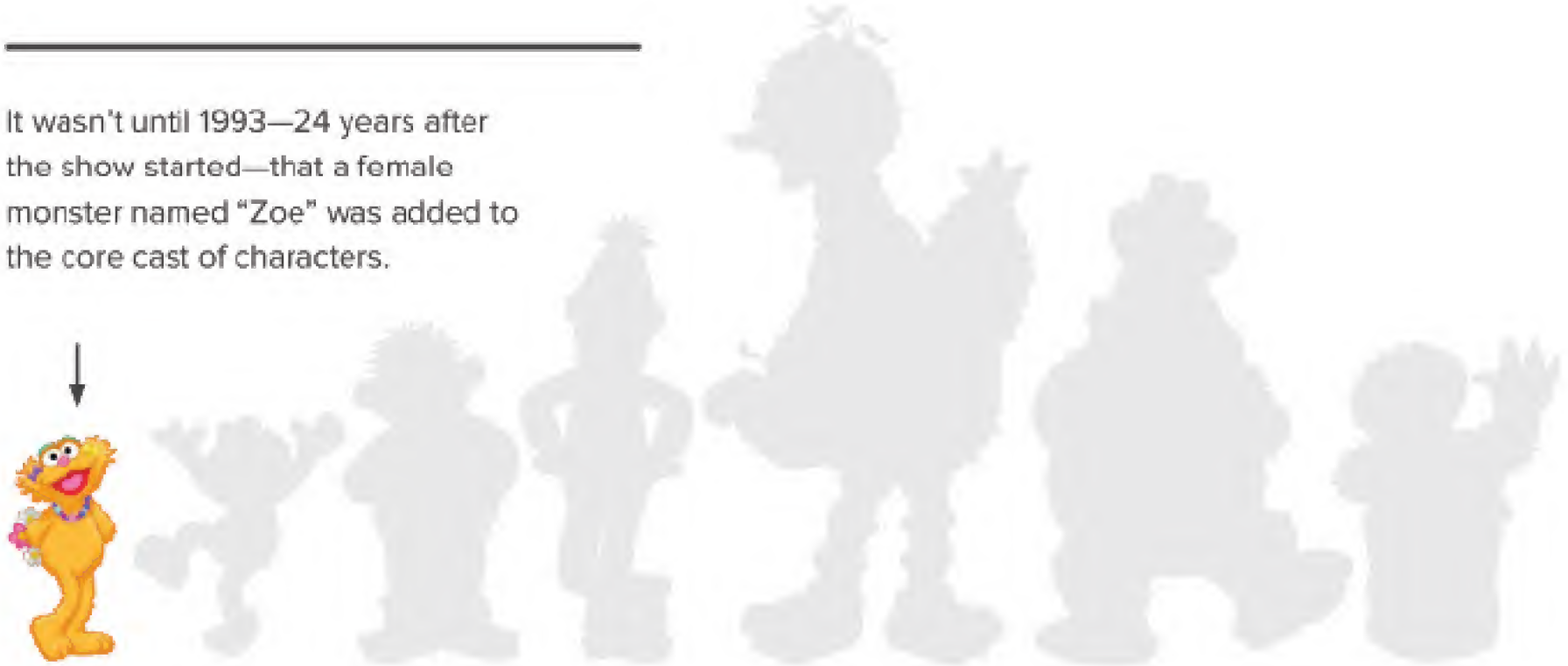
Behind the Scenes: Employment of Women on the Top 100, 250 and 500 Films of 2015, Lauzen

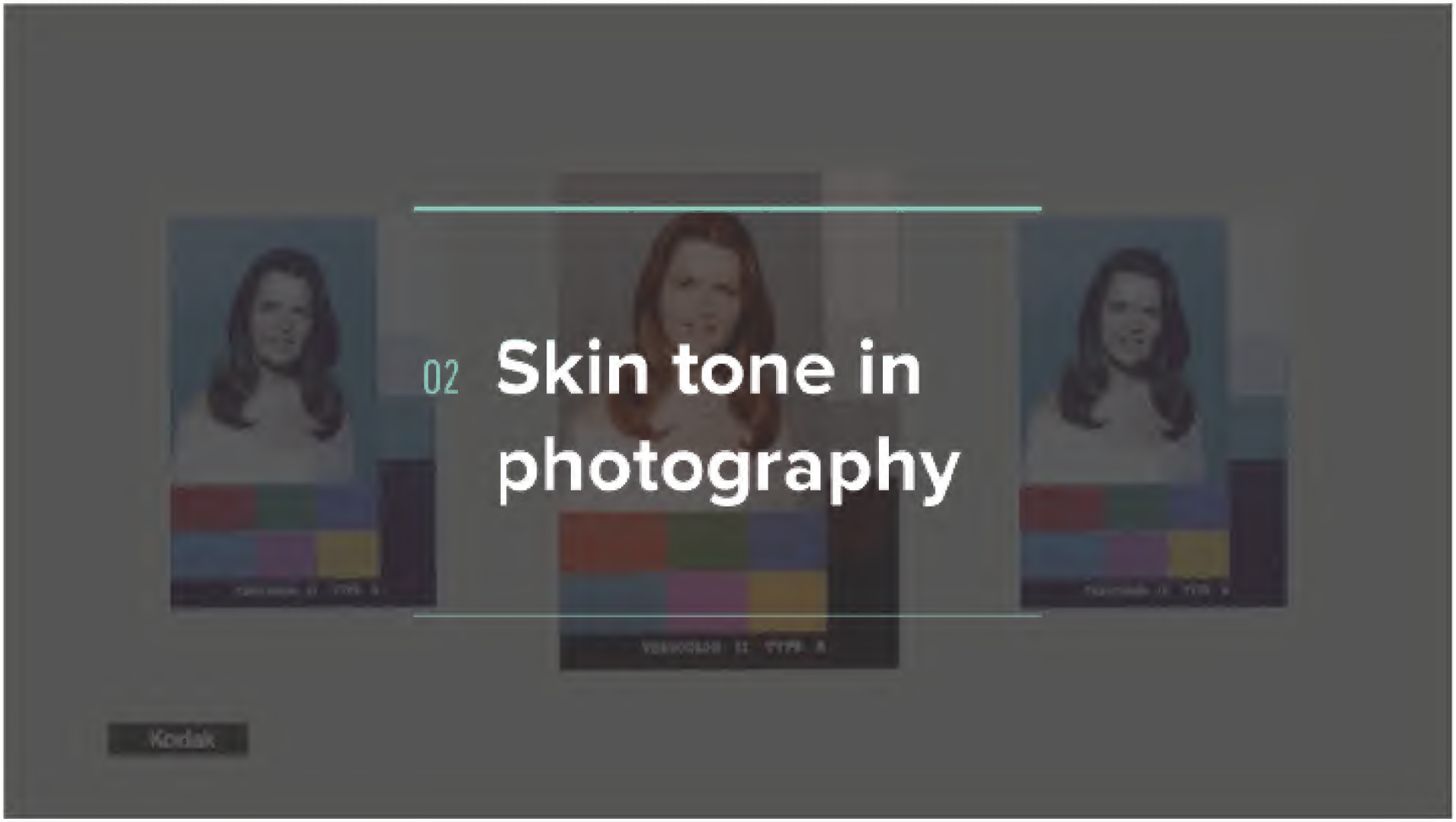
One more observation

Have you ever noticed the “gender” makeup of Sesame Street?

All of the original starring monsters are regarded with male pronouns, despite having no discernable physically gendered traits.

It wasn't until 1993—24 years after the show started—that a female monster named "Zoe" was added to the core cast of characters.




02 Skin tone in photography
Kodak

SKIN TONE IN PHOTOGRAPHY

This is a “Shirley Card”

Named after a Kodak studio model named Shirley Page, they were the primary method for calibrating color when processing film.

SOURCES

Color film was built for white people. Here's what it did to dark skin. (Vox)

How Kodak's Shirley Cards Set Photography's Skin-Tone Standard. NPR

SKIN TONE IN PHOTOGRAPHY

Until about 1990, virtually all Shirley Cards featured caucasian women.

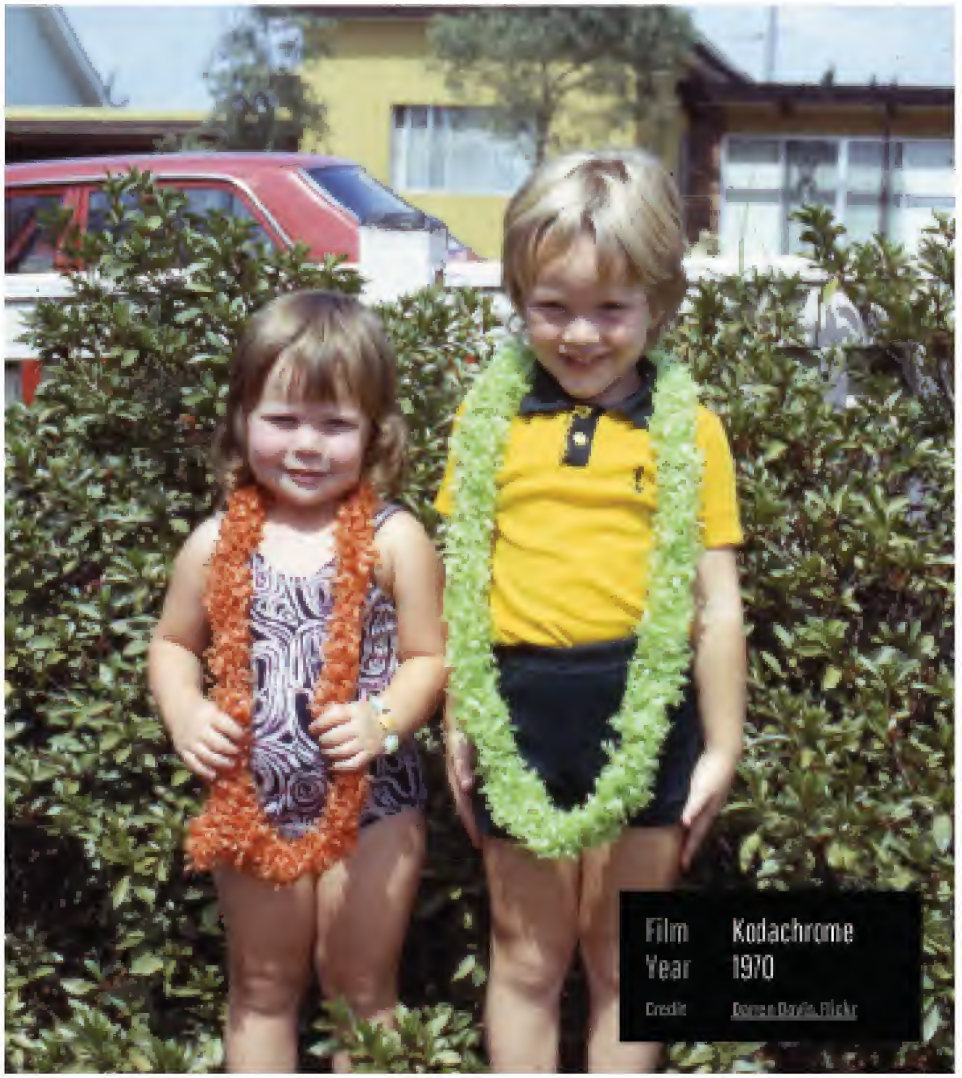
SOURCES

Color film was built for white people. Here's what it did to dark skin. (Vox)

Colour Balance, Image Technologies, and Cognitive Equity. Roll

How Photography Was Optimized for White Skin Color (Priceonomics)

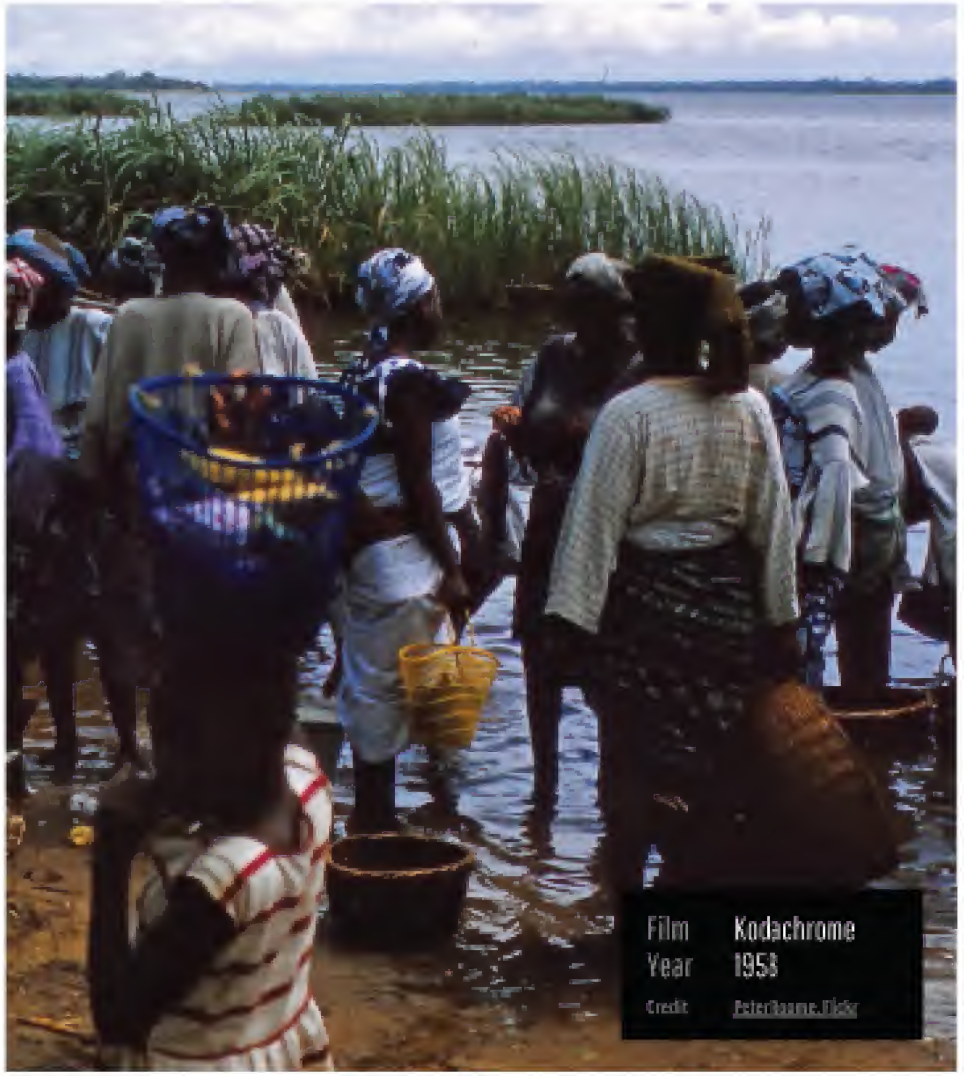
As a result, photos featuring people with light skin looked fairly accurate.



SOURCES

Color film was built for white people. Here's what it did to dark skin. (Vox)
Colour Balance, Image Technologies, and Cognitive Equity, Roth
How Photography Was Optimized for White Skin Color (Priceonomics)

Photos featuring people with darker skin, not so much...



SOURCES

Color film was built for white people. Here's what it did to dark skin. (Vox)
Colour Balance, Image Technologies, and Cognitive Equity. (Roth)
How Photography Was Optimized for White Skin Color (Priceonomics)

And when there was a mix, the difference was most noticeable.

SOURCES

Color film was built for white people. Here's what it did to dark skin. (Vox)
Colour Balance, Image Technologies, and Cognitive Equity, Roth
How Photography Was Optimized for White Skin Color (Priceonomics)

SKIN TONE IN PHOTOGRAPHY

As society became more integrated, photographers found workarounds.

The most common techniques were to shoot with significantly brighter lights (making the room really hot!), using a stronger flash (42% brighter!), and preparing separate cameras with different calibrations for people with different skin tones.

SOURCES

Colour Balance, Image Technologies, and Cognitive Equity, Roth

How Photography Was Optimized for White Skin Color (Priceonomics)

But motivation for innovation came from chocolatiers and wood furniture manufacturers.

Kodak was receiving complaints that they weren't getting the right brown tones on chocolates, and that stains and wood grains were not true to life.

SOURCES

Color film was built for white people. Here's what it did to dark skin. (Vox)

Colour Balance, Image Technologies, and Cognitive Equity, Roth

How Photography Was Optimized for White Skin Color (Priceonomics)

SKIN TONE IN PHOTOGRAPHY

Related research

As of 2014, across all 2 million Implicit Association Test (IAT) participants

51%

showed a moderate to strong bias for white faces

As of 2014, across all U.S. white IAT participants, the median "D score" was

0.402

indicating a moderate bias for white faces

As of 2011, of adults working in Engineering roles in the U.S. (age 16 and over)

4.8%

were black

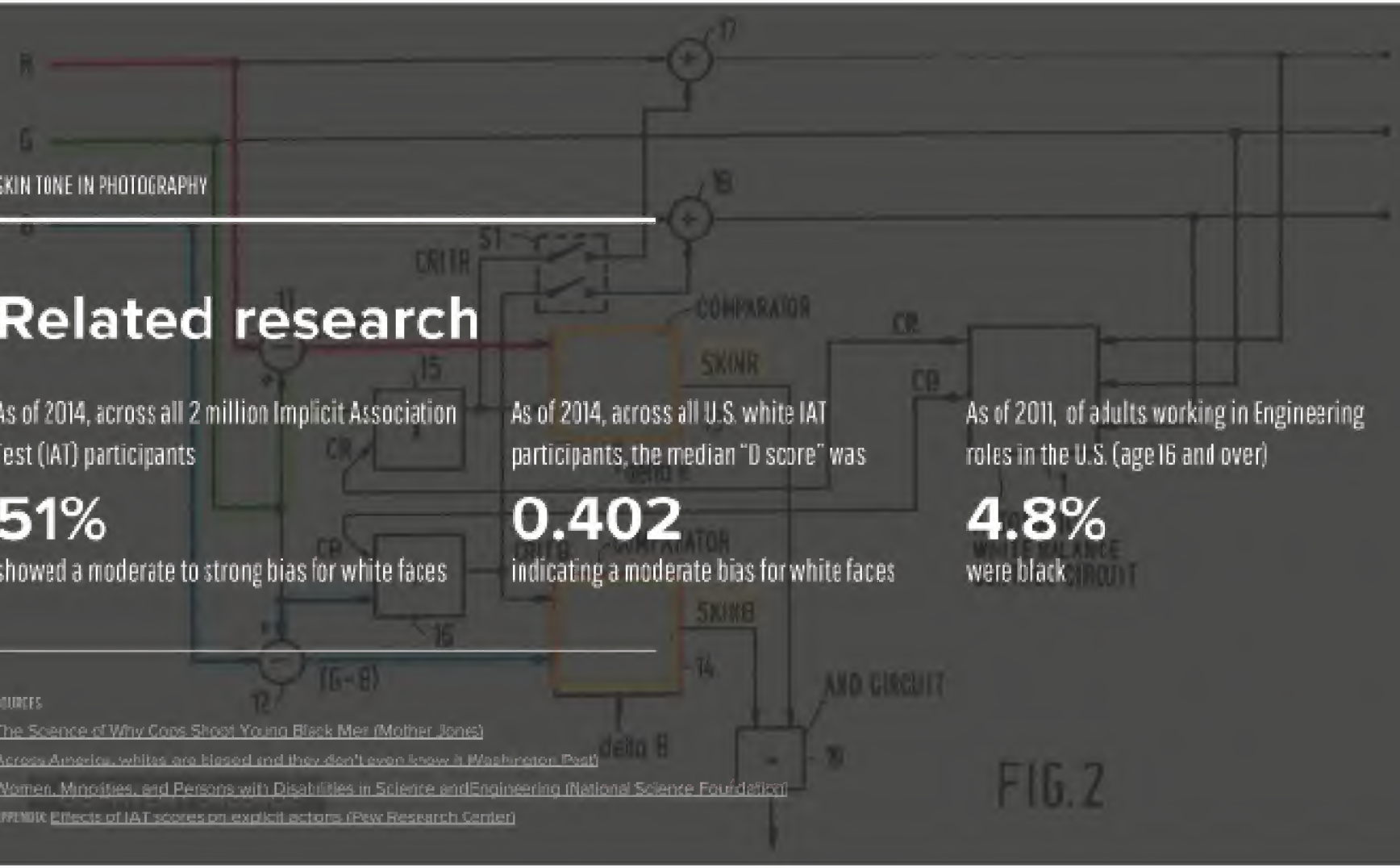


FIG. 2

SOURCES

The Science of Why Cops Shoot Young Black Men (Mother Jones)

Across America, whites are biased and they don't even know it (Washington Post)

Women, Minorities, and Persons with Disabilities in Science and Engineering (National Science Foundation)

APPENDIX: Effects of IAT scores on explicit actions (Pew Research Center)

03 Automotive safety

AUTOMOTIVE SAFETY

Until 2011, female body-type crash test dummies were not required by the United States Department of Transportation.

SOURCE

Female dummy makes her mark on male-dominated crash tests Washington Post

AUTOMOTIVE SAFETY

As a result, female drivers are at a higher risk behind the wheel.

Odds a female driver will sustain severe injuries in an accident

47%

▲ higher than a male driver

SOURCE

Vulnerability of female drivers involved in motor vehicle crashes: an analysis of US population at risk, Bose, Segui-Gomez, and Crandall

AUTOMOTIVE SAFETY

Things are improving, but the target percentile for female test dummies remains problematic.

Male body percentile

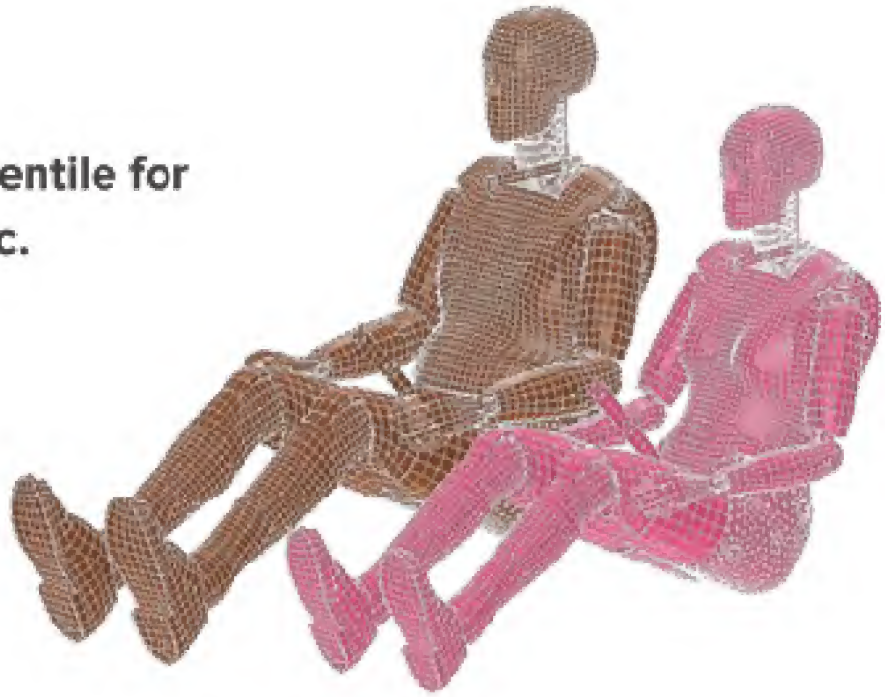
50th

5'9" 176 lbs

Female body percentile

5th

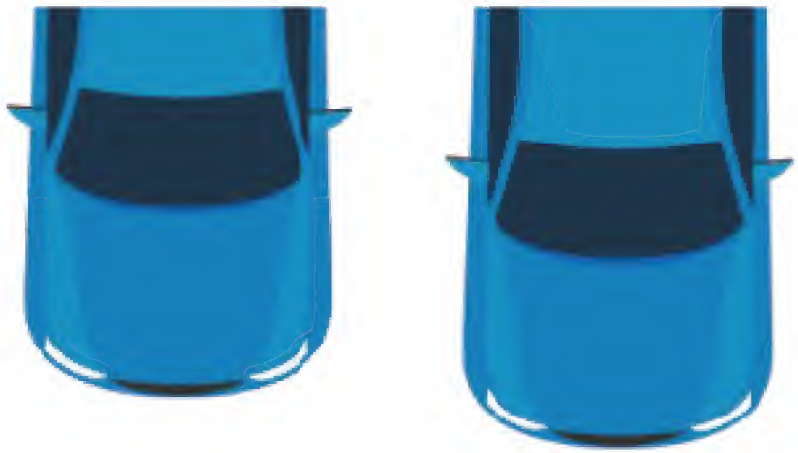
4'11" 108 lbs



SOURCE

Female dummy makes her mark on male-dominated crash tests Washington Post)

Female crash test dummy can reduce injurie (Chalmers)



Related research

Scenarios tested using female body-type crash test dummies in the driver's seat

1 of 3

As of 2015, the number of women working in the motor vehicle manufacturing industry

26.7%

As of 2015, the number of car buying decisions influenced by women

80%

SOURCE

Women in Cars - A Mega Trend for the Automotive Industry (Frost & Sullivan)

Vehicle safety ratings (National Highway Traffic Safety Administration)

Labor Force Statistics (United States Department of Labor)

We can't remove human perception from the loop.

And we can't be gripped by inaction either.

The inequity demonstrated in these examples may feel overwhelming, perhaps even a little disheartening.

But we're in the right place at the right time and in the right industry to do something about it.